

# Detecting Damped Lyman- $\alpha$ Absorbers with Gaussian Processes

Roman Garnett,<sup>1</sup><sup>★</sup> Shirley Ho,<sup>2</sup> Simeon Bird,<sup>3</sup> and Jeff Schneider<sup>4</sup>

<sup>1</sup>*Department of Computer Science and Engineering, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130, USA*

<sup>2</sup>*Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

<sup>3</sup>*Department of Physics & Astronomy, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

<sup>4</sup>*School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA*

17 May 2016

## ABSTRACT

We develop an automated technique for detecting damped Lyman- $\alpha$  absorbers (DLAs) along spectroscopic sightlines to quasi-stellar objects (QSOs or quasars). The detection of DLAs in large-scale spectroscopic surveys such as SDSS-III sheds light on galaxy formation at high redshift, showing the nucleation of galaxies from diffuse gas. We use nearly 50 000 QSO spectra to learn a novel tailored Gaussian process model for quasar emission spectra, which we apply to the DLA detection problem via Bayesian model selection. We propose models for identifying an arbitrary number of DLAs along a given line of sight. We demonstrate our method’s effectiveness using a large-scale validation experiment, with excellent performance. We also provide a catalog of our results applied to 162 861 spectra from SDSS-III data release 12.

## 1 INTRODUCTION

The damped Lyman- $\alpha$  (Ly $\alpha$ ) systems (DLAs) (Wolfe et al. 1986, 2005) define the class of absorption-line systems discovered in the restframe UV spectra of distant quasars, with H I column densities  $N_{\text{H I}} > 2 \times 10^{20} \text{ cm}^{-2}$ , as measured from the analysis of damping wings in the Ly $\alpha$  profile. Recent spectroscopic quasar surveys such as the Sloan Digital Sky Survey (SDSS) (York et al. 2000) have produced a vast sample of quasar spectra showing Ly $\alpha$  absorption at  $z > 2$ . SDSS-III has measured nearly 300 000 quasar spectra over its brief history. Even larger surveys, such as SDSS-IV and DES<sup>1</sup>, soon plan to observe 1–2 million quasars. Finding DLAs in these surveys has historically involved a combination of automated template fitting and visual inspection, but this is clearly infeasible with the size of upcoming datasets. Furthermore, SDSS data trades low signal-to-noise ratios for statistical power, making detection of even distinctive signals such as DLAs substantially harder, and making noise-induced systematic error hard to control.

There have been several previous DLA searches in SDSS. These include a visual-inspection survey (Slosar et al. 2011), visually guided Voigt-profile fitting (Prochaska et al. 2005; Prochaska & Wolfe 2009) and two automated approaches: a template-matching approach (Noterdaeme et al. 2012), and an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012). While these methods have had some success in detecting large DLA catalogs, their manual components made them labor intensive, and their reliance on templates made them subject to hard-to-quantify systematic biases.

In this paper, we present a new, completely automated method based on a rigorous Bayesian model-selection framework. We model the quasar spectra, including the continuum and non-DLA absorption, using Gaussian process (Rasmussen & Williams 2006) models with a bespoke covariance function. Earlier catalogs are used as prior

information to train the covariance. We provide a catalog of our results on 162 861 QSOs with  $z \geq 2.15$  from data release 12 of SDSS-III, demonstrating that our method scales to very large datasets, making it ideally suited for future surveys. Furthermore, as our method relies on a well-defined probabilistic framework, it allows us to estimate the probability that each system is indeed a DLA, rather than a noise fluctuation, degrading gracefully for low signal-to-noise observations. This property allows us to obtain substantially more-reliable measurements of the statistics of the DLA population (Bird et. al, in preparation) and to extend our catalog to high redshift even with low-quality data.

Our method is applicable not just to DLAs, but also to other classes of absorption systems, such as Lyman limit systems and metal absorbers, which we intend to examine in future work. We focus on DLAs here both because of the large body of prior work which enables us to thoroughly verify our catalogs, and the intrinsic importance of these systems.

DLAs are a direct probe of neutral gas at densities close to those required to form stars (Cen 2012). They thus provide a powerful independent check on models of galaxy formation, allowing us to directly probe non-luminous material at high redshift. The exact nature of the systems hosting DLAs was initially debated, with kinematic data combined with simple semi-analytic models appearing to indicate objects similar in size to present day star-forming galaxies (Prochaska & Wolfe 1997; Jedamzik & Prochaska 1998; Maller et al. 2001), whereas early simulations produced clumps closer in size to dwarf galaxies (Haehnelt et al. 1998; Okoshi & Nagashima 2005). Recent numerical simulations are able to reproduce most observations with neutral hydrogen clouds stretching almost to the virial radius of objects larger than dwarfs, but smaller than present day star-forming galaxies (Pontzen et al. 2008; Rahmati et al. 2013; Bird et al. 2015). Associated galactic stellar components have been detected in a few, particularly neutral hydrogen and metal-rich systems at low redshift (Le Brun et al. 1997; Rao et al. 2003; Chen 2005). However, unbiased surveys have placed strong upper lim-

<sup>1</sup> <http://desi.lbl.gov>

its on the star-formation rates of the median DLA (Fumagalli et al. 2015), indicating that DLAs are associated with low star-formation rate objects.

DLAs currently represent our only probe of small- to moderate-sized galaxies at high redshift, and are known to have dominated the neutral-gas content of the Universe from redshift  $z = 5$  (when the Universe was 1.2 Gyr old) to today (Gardner et al. 1997; Wolfe et al. 2005). These systems likely played a significant role in fuelling star formation across the cosmic time, and thus their abundance as a function of redshift provides strong constraints on models of galaxy formation (Bird et al. 2014). Our work, including publicly available software, will not only provide observers with a new automated tool for detecting these objects, but also provide theorists with a reliable catalog on which to base theoretical models.

## 2 NOTATION

We will briefly establish some notation. Consider a qso with redshift  $z_{\text{qso}}$ ; we will always assume that  $z_{\text{qso}}$  is known, allowing us to work in the quasar restframe. We will notate a qso's true emission spectrum by a function  $f: \mathbb{R} \rightarrow \mathbb{R}$ , where  $f(\lambda)$  represents the flux corresponding to rest wavelength  $\lambda$ . Without subscript,  $\lambda$  will always refer to quasar rest wavelengths,  $\lambda_{\text{rest}}$ , rather than observed wavelengths,  $\lambda_{\text{obs}}$ . Note that the spectral emission function  $f$  is never directly observed, both due to measurement error and due to absorption by intervening matter along the line of sight. We will denote the observed flux by a corresponding function  $y(\lambda)$ , which will again be a function of the rest wavelengths.

Spectroscopic observations of a qso are made at a discrete set of wavelengths  $\lambda$ , for which we observe a corresponding vector of flux measurements  $\mathbf{y}$ , where we have defined  $y_i = y(\lambda_i)$ . For a given qso, we will represent the set of observation locations and values  $(\lambda, \mathbf{y})$  by  $\mathcal{D}$ .

We will often encounter data with missing values due to observation-dependent pixel masking. When required, we will represent these in the text with a special value called NaN (for “not a number”). Calculations on data containing NaNs will always ignore these values.

## 3 BAYESIAN MODEL SELECTION

Our approach to DLA detection will depend on *Bayesian model selection*, which will allow us to directly compute the probability that a given quasar sightline contains a DLA. We will develop two probabilistic models for a given set of spectroscopic observations  $\mathcal{D}$ : one for sightlines with intervening DLAs, and one for those without. Then, given the available data, we will compute the posterior probability that the former model is correct. We will give a high-level overview of Bayesian model selection below, then proceed to describe our models for DLA detection below.

Let  $\mathcal{M}$  be a probabilistic model, and let  $\theta$  represent a vector of parameters for this model (if any). Given a set of observed data  $\mathcal{D}$  and a set of candidate models  $\{\mathcal{M}_i\}$  containing  $\mathcal{M}$ , we wish to compute the probability of  $\mathcal{M}$  being the correct model to explain  $\mathcal{D}$ . The key quantity of interest to model selection is the so-called *model evidence*:

$$p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \mathcal{M}, \theta) p(\theta | \mathcal{M}) d\theta, \quad (1)$$

which represents the probability of having generating the observed data with the model, after having integrated out any uncertainty in

the parameter vector  $\theta$ . Given the model evidence, we can apply Bayes' rule to compute the posterior probability of the model given the data:

$$\Pr(\mathcal{M} | \mathcal{D}) = \frac{p(\mathcal{D} | \mathcal{M}) \Pr(\mathcal{M})}{p(\mathcal{D})} = \frac{p(\mathcal{D} | \mathcal{M}) \Pr(\mathcal{M})}{\sum_i p(\mathcal{D} | \mathcal{M}_i) \Pr(\mathcal{M}_i)}, \quad (2)$$

where  $\Pr(\mathcal{M})$  represents the prior probability of the model. Notice that computing the posterior probability of  $\mathcal{M}$  requires computing the normalizing constant in the denominator, which requires computing the model evidence of every model being considered.

We will develop two models for spectroscopic observations of qsos,  $\mathcal{M}_{\text{noDLA}}$ , for lines of sight that do not contain intervening DLAs, and  $\mathcal{M}_{\text{DLA}}$ , for those that do. Both of these models will rely heavily on Gaussian processes, which we will introduce below.

## 4 GAUSSIAN PROCESSES

The main object of interest we wish to perform inference about is a given qso's emission function  $f(\lambda)$ . This is in general a complicated function with no simple parametric form available, so we will instead use nonparametric inference techniques to reason about it. *Gaussian processes* (GPs) provide a powerful nonparametric framework for modeling unknown functions, which we will adopt for this task. See Rasmussen & Williams (2006) for an extensive introduction to GPs.

### 4.1 Definition and prior distribution

Let  $\mathcal{X}$  be an arbitrary input space, for example the real line  $\mathbb{R}$ , and let  $f: \mathcal{X} \rightarrow \mathbb{R}$  be a real-valued function on  $\mathcal{X}$  we wish to model. We will continue to use  $\lambda$  to indicate inputs to the function  $f$ . A Gaussian process is an extension of the multivariate Gaussian distribution  $\mathcal{N}(\mu, \Sigma)$  to infinite domains. Like the multivariate Gaussian distribution, a GP is fully specified by its first two central moments: a mean function  $\mu(\lambda)$  and a positive semidefinite covariance function  $K(\lambda, \lambda')$ .<sup>2</sup>

$$\mu(\lambda) = \mathbb{E}[f(\lambda) | \lambda]; \quad (3)$$

$$K(\lambda, \lambda') = \text{cov}[f(\lambda), f(\lambda') | \lambda, \lambda']. \quad (4)$$

The former describes the pointwise expected value of the function and the latter describes the correlation structure around the mean. Given  $\mu$  and  $K$ , we may endow the function space  $f$  with a Gaussian process prior probability distribution:

$$p(f) = \mathcal{GP}(f; \mu, K). \quad (5)$$

The defining characteristic of a Gaussian process is that given a finite set of inputs  $\lambda$ , the corresponding vector of function values  $\mathbf{f} = f(\lambda)$  is multivariate Gaussian distributed:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}; \mu(\lambda), K(\lambda, \lambda)), \quad (6)$$

where the mean vector and covariance matrix are derived simply by evaluating the mean and covariance functions at the inputs  $\lambda$ , and the multivariate Gaussian probability distribution function is given by

$$\mathcal{N}(\mathbf{f}; \mu, \mathbf{K}) = \frac{1}{\sqrt{(2\pi)^d \det \mathbf{K}}} \exp\left(-\frac{1}{2}(\mathbf{f} - \mu)^\top \mathbf{K}^{-1}(\mathbf{f} - \mu)\right), \quad (7)$$

where  $d$  is the dimension of  $\mathbf{f}$ .

<sup>2</sup> A function  $K: \mathcal{X}^2 \rightarrow \mathbb{R}$  is *positive semidefinite* if, for every finite subset  $\Lambda = \{\lambda_i\}_{i=1}^n \subset \mathcal{X}$ , the  $n \times n$  Gram matrix  $\mathbf{A}$ , defined by  $A_{ij} = K(\lambda_i, \lambda_j)$ , satisfies  $\mathbf{c}^\top \mathbf{A} \mathbf{c} \geq 0$  for all  $\mathbf{c} \in \mathbb{R}^n$ .

## 4.2 Conditioning on observations

Suppose that we have a function of interest  $f$  and have chosen a GP prior for  $f$  as above:

$$p(f) = \mathcal{GP}(f; \mu, K). \quad (8)$$

Given a set of (potentially noisy) observations of the function  $\mathcal{D}$ , we may update our belief about  $f$  to reflect the information contained in these observations. We may then use this posterior distribution to reason further about  $f$ .

### 4.2.1 Observation model

Consider a set of noisy observations  $\mathcal{D} = (\lambda, \mathbf{y})$  made at input locations  $\lambda$ . Our Gaussian process prior on  $f$  implies a multivariate Gaussian distribution for the corresponding (unknown, so-called *latent*) function values  $\mathbf{f} = f(\lambda)$ , but does not specify the relationship between these values and our observations  $\mathbf{y}$ . Instead we must further model the mechanism generating our observations, which we will encode by a distribution

$$p(\mathbf{y} | \lambda, \mathbf{f}). \quad (9)$$

In general this can be any arbitrary probabilistic model, but here we will assume additive Gaussian noise.

Given a single input location  $\lambda$ , we assume that the corresponding observed value  $y$  is realized by corrupting the true value of the latent function  $f(\lambda)$  by zero-mean additive Gaussian noise with known variance  $\sigma(\lambda)^2$ :

$$p(y | \lambda, f(\lambda), \sigma(\lambda)) = \mathcal{N}(y; f(\lambda), \sigma(\lambda)^2). \quad (10)$$

We assume the noise process is independent for every  $\lambda$ , but note that we do not make a homoskedasticity assumption; rather, we allow the noise variance to depend on  $\lambda$ . This capability to handle heteroskedastic noise is critical for the analysis of spectroscopic measurements, where the noise associated with flux measurements can vary widely as a function of wavelength.

Returning to our entire set of observations  $\mathcal{D} = (\lambda, \mathbf{y})$ , we assume that the noise variance associated with each of these measurements is known and given by a corresponding vector  $\mathbf{v}$ , with  $v_i = \sigma(\lambda_i)^2$ . Given our model for individual observations (10) and the noise independence assumption, the entire observation model is given by

$$p(\mathbf{y} | \lambda, \mathbf{f}, \mathbf{v}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{V}), \quad (11)$$

where  $\mathbf{V} = \text{diag } \mathbf{v}$ .

### 4.2.2 Prior of noisy observations

Given a set of observations locations  $\lambda$  and a corresponding vector of noise variances  $\mathbf{v}$ , we may use the above to compute the prior distribution for a corresponding vector of observations  $\mathbf{y}$  by marginalizing the vector of latent function values  $\mathbf{f}$ :

$$\begin{aligned} p(\mathbf{y} | \lambda, \mathbf{v}) &= \int p(\mathbf{y} | \lambda, \mathbf{f}, \mathbf{v}) p(\mathbf{f} | \lambda) d\mathbf{f} \\ &= \int \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{V}) \mathcal{N}(\mathbf{f}; \mu(\lambda), K(\lambda, \lambda)) d\mathbf{f} \\ &= \mathcal{N}(\mathbf{y}; \mu(\lambda), K(\lambda, \lambda) + \mathbf{V}), \end{aligned} \quad (12)$$

where we have used the fact that Gaussians are closed under convolution to compute the integral in closed form.

### 4.2.3 Deriving the posterior distribution

Now let  $\lambda^*$  be an arbitrary finite set of additional input locations, and let  $\mathbf{f}^* = f(\lambda^*)$  represent the vector of corresponding function values. We may use (6) and (12) to derive the joint prior distribution of  $\mathbf{y}$  and  $\mathbf{f}^*$ :

$$p\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \middle| \lambda, \lambda^*, \mathbf{v}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix}; \begin{bmatrix} \mu(\lambda) \\ \mu(\lambda^*) \end{bmatrix}, \begin{bmatrix} K(\lambda, \lambda) + \mathbf{V} & K(\lambda, \lambda^*) \\ K(\lambda^*, \lambda) & K(\lambda^*, \lambda^*) \end{bmatrix}\right). \quad (13)$$

Now we may use a standard Gaussian identity to derive the posterior distribution of  $\mathbf{f}^*$  conditioned on the observed values  $\mathbf{y}$ :

$$p(\mathbf{f}^* | \lambda^*, \mathcal{D}, \mathbf{v}) = \mathcal{N}(\mathbf{f}^*; \mu_{f|\mathcal{D}}(\lambda^*), K_{f|\mathcal{D}}(\lambda^*, \lambda^*)), \quad (14)$$

where

$$\mu_{f|\mathcal{D}}(\lambda) = \mu(\lambda) + K(\lambda, \lambda)(K(\lambda, \lambda) + \mathbf{V})^{-1}(\mathbf{y} - \mu(\lambda)); \quad (15)$$

$$K_{f|\mathcal{D}}(\lambda, \lambda') = K(\lambda, \lambda') - K(\lambda, \lambda)(K(\lambda, \lambda) + \mathbf{V})^{-1}K(\lambda, \lambda'). \quad (16)$$

We recognize that the posterior is a Gaussian process with mean function  $\mu_{f|\mathcal{D}}$  and covariance function  $K_{f|\mathcal{D}}$  that have been updated to reflect the information contained in the observations. This closure under conditioning and the ability to compute the posterior analytically under the Gaussian noise assumption are two convenient properties of GPs.

In typical applications of GP inference, the prior mean function  $\mu$  and prior covariance function  $K$  would be selected from numerous several off-the-shelf solutions available for this purpose; however, none of these would be directly appropriate for modeling qso emission spectra, due to their somewhat complex nature. Typical parametric covariance functions, for example, tend to be translation invariant and encode strictly decreasing covariance as a function of the distance between inputs.<sup>3</sup> qso emission spectra, however, are neither stationary, nor should we expect the covariance to be diagonal dominant. For example, strong off-diagonal correlations must exist between potentially distant emission lines, such as members of the Lyman series. Rather, below we will construct a custom GP prior distribution for modeling these spectra in the next section.

## 5 LEARNING A GP PRIOR FOR QSO SPECTRA

We wish to construct a Gaussian process prior for qso spectra, specifically, those that do not contain an intervening DLA along the line of sight. This will form the basis for our null model  $\mathcal{M}_{\text{-DLA}}$ . We will later extend this to form our DLA model  $\mathcal{M}_{\text{DLA}}$ .

As described in the previous section, a Gaussian process is defined entirely by its first two moments: a mean function  $\mu(\lambda)$  and a covariance function  $K(\lambda, \lambda')$ . Therefore, our goal in this section will be to derive reasonable prior choices for these functions. Due to the complex structure of qso emission spectra, our approach will be to make as few assumptions as possible. Instead, we adopt a data-driven approach and learn an appropriate model given over 48 000 examples contained in a previously compiled catalog of quasar spectra recorded by the BOSS spectrograph (Smee et al. 2013). We describe the data below.

<sup>3</sup> The Wiener process, modeling the sample paths associated with Brownian motion, is a Gaussian process with such a covariance function.

### 5.1 Data

Together, sdss-I, -II (Abazajian et al. 2009), and -III (Eisenstein et al. 2011) used a drift-scanning mosaic CCD camera (Gunn et al. 1998) to image over one-third of the sky (14 555 square degrees) in five photometric bandpasses (Fukugita et al. 1996; Smith et al. 2002; Doi et al. 2010) to a limiting magnitude of  $r < 22.5$  using the dedicated 2.5 m Sloan Telescope (Gunn et al. 2006) located at Apache Point Observatory in New Mexico.

The Baryon Oscillation Spectroscopic Survey (BOSS), a part of the sdss-III survey (Eisenstein et al. 2011) has obtained spectra of 1.5 million galaxies approximately volume limited out to  $z \sim 0.6$  (Reid et al. 2016), and an additional 150 000 spectra of high-redshift quasars and ancillary sources. BOSS has measured the characteristic scale imprinted by baryon acoustic oscillations (BAOs) in the early Universe from the spatial distribution of galaxies at  $z \sim 0.5$  and the H I absorption lines in the intergalactic medium at  $z \sim 2.3$  (Anderson et al. 2012, 2014; Aubourg et al. 2015). The quasar target selection is described in (Ross et al. 2012; Bovy et al. 2011). Here we use data included in data releases 9 (DR9) (Ahn et al. 2012) and 12 (DR12) (Ahn et al. 2014) of sdss-III; in particular, we primarily use the associated quasar catalogs from various data releases<sup>4</sup> (Pâris et al. 2012, 2014).

#### 5.1.1 Description of data

We used the qso spectra from the BOSS DR9 Lyman- $\alpha$  forest sample (Lee et al. 2013) to train our GP model. This sample comprises 54 468 qso spectra with  $z_{\text{qso}} > 2.15$  from the DR9 release appropriate for Lyman- $\alpha$  forest analysis. An analogous model built from the entire DR12 sample will be published along with manuscript for general-purpose use, along with the source code (in MATLAB) we used to train our model.

The Lyman- $\alpha$  forest sample was augmented with a previously compiled “concordance” DLA catalog (Carithers 2012), combining the results of three previous DLA searches. These include a visual-inspection survey (Slosar et al. 2011) and two previous automated approaches: a template-matching approach (Noterdaeme et al. 2012), and an unpublished machine-learning approach using Fisher discriminant analysis (Carithers 2012). Any line of sight flagged in at least two of these catalogs as containing a DLA is included in the concordance catalog. Both previous automated DLA searches also produced estimates of the absorber redshift  $z_{\text{DLA}}$  and column density  $\log_{10} N_{\text{H I}}$ . The concordance catalog also includes these estimates for flagged sightlines; when a sightline is included in both automated catalogs, the arithmetic mean of the associated estimates was recorded. A total of 5 854 lines of sight are flagged as containing an intervening DLA in the catalog (10.7%).

### 5.2 Modeling decisions

To avoid effects due to redshift, we will build our emission model for wavelengths in the rest frame of the qso. Furthermore, to account for arbitrary scaling of flux measurements, we will build a GP prior for normalized flux. Specifically, given the observed coadded flux of a qso, we normalize all flux measurements by dividing by the median flux observed between 1270 Å and 1290 Å in the rest frame of the

qso, a region redwards of the Ly $\alpha$  forest and void of major emission features.

Because this study is concerned with identifying DLAs, we will only model the flux bluewards of the Ly $\alpha$  emission in the rest frame of a given qso.<sup>5</sup> Specifically, we model emissions in the range spanning from the Lyman limit to the Lyman- $\alpha$  line in the qso restframe.<sup>6</sup> Our approach will be to learn a mean vector and covariance matrix on a dense grid of wavelengths in this range, which we will then interpolate as required by a particular set of observed wavelengths. The chosen grid was the set of wavelengths

$$\lambda \in [911.75 \text{ Å}, 1216.75 \text{ Å}], \quad (17)$$

with a linearly equal spacing of  $\Delta\lambda = 0.25 \text{ Å}$ .<sup>7</sup> This resulted in a vector of input locations  $\lambda$  with  $|\lambda| = N_{\text{pixels}} = 1\,221$  pixels.

Given a GP prior for qso emission spectra,  $p(f) = \mathcal{GP}(f; \mu, K)$ , the prior distribution for emissions on the chosen grid  $\lambda$ ,  $\mathbf{f} = f(\lambda)$  is a multivariate Gaussian:

$$p(\mathbf{f} | \lambda, z_{\text{qso}}) = \mathcal{N}(\mathbf{f}; \mu, \mathbf{K}), \quad (18)$$

where  $\mu = \mu(\lambda)$  and  $\mathbf{K} = K(\lambda, \lambda)$ . Note that we must condition on the qso redshift  $z_{\text{qso}}$  because it is required for shifting into the quasar restframe.

As mentioned previously, however, we can never observe  $f$  directly, both due to measurement error and due to absorption by intervening matter along the line of sight. The former can be handled easily for our spectra by using the pipeline error estimates in the role of the noise vector  $\mathbf{v}$  (see Section 4.2.1). However, the latter is more problematic, especially in our chosen region, which includes the Lyman- $\alpha$  forest. In principle, if we knew the exact nature of the intervening matter, we could model this absorption explicitly; however, this is unrealistic. We will instead model the effect of small absorption phenomena (absorption by objects with column density below the DLA limit,  $\log_{10} N_{\text{H I}} < 20.3$ ) by an additional additive wavelength-dependent Gaussian noise term, which we will learn. Therefore the characteristic “dips” of the Lyman- $\alpha$  forest will be modeled as noisy deviations from the true underlying smooth continuum. Later we will explicitly model larger absorption phenomena (DLAs with  $\log_{10} N_{\text{H I}} \geq 20.3$ ) to build our DLA model  $\mathcal{M}_{\text{DLA}}$ .

The mathematical consequence of this modeling decision is as follows. Consider the arbitrary GP model in (18). We wish to model the associated spectroscopic observation values on the chosen grid,  $\mathbf{y} = y(\lambda)$ . Suppose that the measurement noise vector  $\mathbf{v} = \sigma(\lambda)^2$  has been specified. During our exposition on GPs, we described the additive Gaussian noise observation model

$$p(\mathbf{y} | \mathbf{f}, \mathbf{v}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{V}). \quad (19)$$

The model we adapt here will involve a shared non-DLA absorption “noise” vector  $\omega$  modeling absorption deviations from the qso

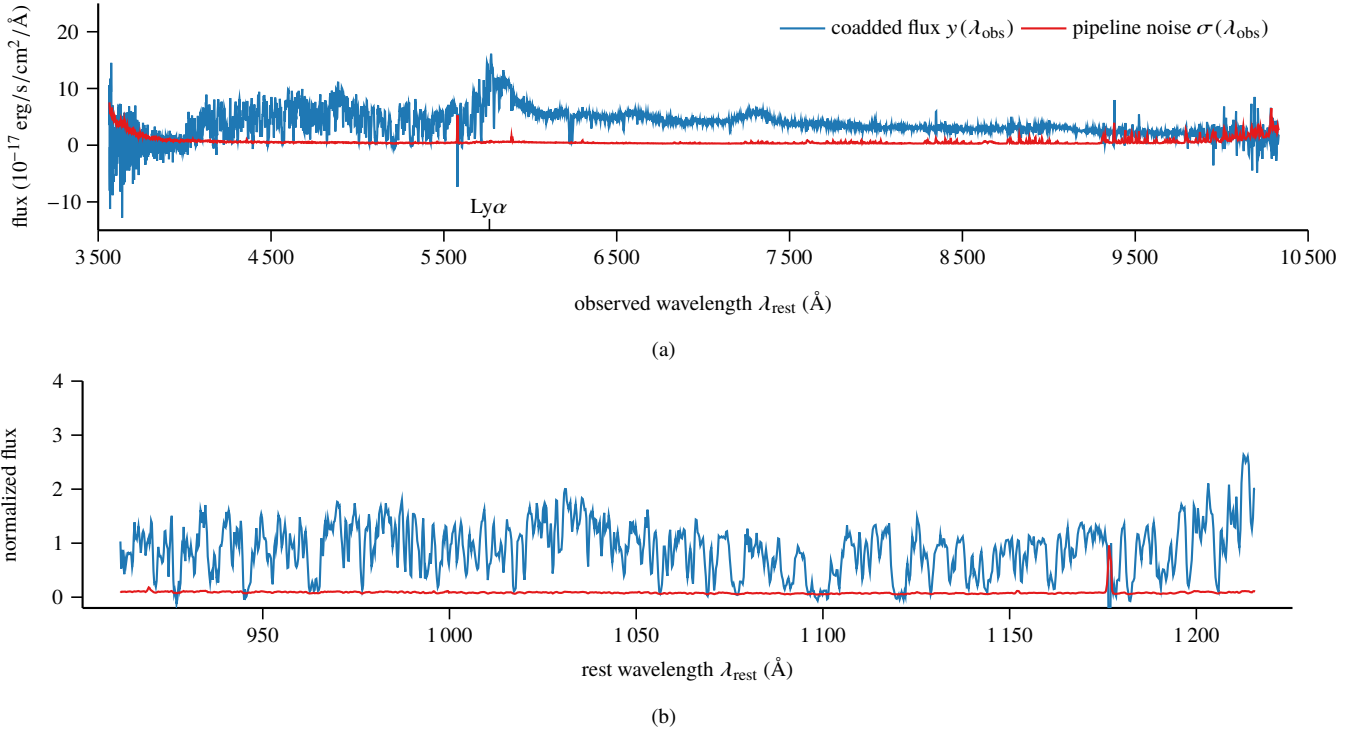
<sup>4</sup> <http://www.sdss.org/dr12/algorithms/booss-dr12-quasar-catalog/>

<sup>5</sup> One could consider an extension of our approach where metal absorption lines corresponding to wavelengths redwards of Ly $\alpha$  were considered, requiring modeling spectra over a larger range of wavelengths; however, we will not do so here.

<sup>6</sup> We stop at the Lyman limit to avoid being confused by the Lyman break associated with  $z > 3.5$  Lyman Limit Systems present in the SDSS data (Prochaska et al. 2010).

<sup>7</sup> This represents about 3–4 times the maximum resolution of the BOSS spectrograph; the minimum separation in a single BOSS spectrum’s measured wavelengths is approximately  $(10^{\log_{10} 3600+0.0001} - 3600) \text{ Å} \approx 0.83 \text{ Å}$ . Note, however, that we have tens of thousands of observations corresponding to each of the wavelengths in our chosen grid.





**Figure 1.** An illustration of the data preprocessing procedure for object sdss 020712.80+052753.4, (plate, MJD, fiber) = (4401, 55510, 338);  $z_{\text{qso}} = 3.741$ . This qso is included in the DLA concordance catalog with  $(z_{\text{DLA}}, \log_{10} N_{\text{H}}) = (3.283, 20.39)$ , corresponding to central absorption wavelength  $\lambda_{\text{obs}} = 5206 \text{ \AA}$  or  $\lambda_{\text{rest}} = 1098 \text{ \AA}$  in the qso restframe. The wavelengths are shifted to the qso restframe and pixels outside  $\lambda_{\text{rest}} \in [911.75 \text{ \AA}, 1216.75 \text{ \AA}]$  are discarded. Finally, the flux and noise estimates are normalized by dividing by the median flux in this region. The final result is shown in (b).

continuum. The resulting observation model is

$$p(\mathbf{y} | \mathbf{f}, \mathbf{v}, \omega, z_{\text{qso}}, \mathcal{M}_{\text{-DLA}}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{\Omega} + \mathbf{V}), \quad (20)$$

where  $\mathbf{\Omega} = \text{diag } \omega$ . Therefore, given our chosen input grid  $\lambda$ , the prior distribution of associated spectroscopic observations  $\mathbf{y}$  is

$$p(\mathbf{y} | \mathbf{v}, \omega, z_{\text{qso}}, \mathcal{M}_{\text{-DLA}}) = \mathcal{N}(\mathbf{y}; \mu, \mathbf{K} + \mathbf{\Omega} + \mathbf{V}), \quad (21)$$

derived analogously to (12). Our goal now is to learn appropriate values for  $\mu$ ,  $\mathbf{K}$ , and  $\omega$ , which will fully specify our null model  $\mathcal{M}_{\text{-DLA}}$ .

### 5.3 Learning appropriate parameters

To build our null model, we took the  $N_{\text{spec}} = 48614$  spectra from the BOSS DR9 Lyman- $\alpha$  forest sample that are putatively absent of intervening DLAs. We prepared each of these spectra for processing in an identical manner as follows.

- The augmented spectrum file was loaded and the (wavelength, observed flux, pipeline noise variance) =  $(\lambda, y, v)$  measurements in the chosen modeled region were extracted.
- The wavelengths were shifted to the rest frame of the qso.
- Flux measurements with serious pixel mask bit flags (FULLREJECT, NOSKY, BRIGHTSKY, NODATA) set by the sdss pipeline were masked (replaced by NaN).
- The flux normalizer was determined by examining the region corresponding to  $[1270, 1290] \text{ \AA}$  in the restframe of the quasar; the median nonmasked value in this range was used for normalization.
- The flux and noise variance were normalized with the value computed in the last step.

Finally, we linearly interpolated the resulting flux and noise variance measurements of each spectrum onto the chosen wavelength grid  $\lambda$ . Note that this interpolation preserved NaNs; we did not “interpolate through” masked pixels. We also did not extrapolate beyond the range of wavelengths present in each spectrum. The preprocessing procedure is illustrated in Figure 1 on a spectrum we will use as a running example.

We collect the resulting interpolated vectors into  $(N_{\text{spec}} \times N_{\text{pixels}})$  matrices  $\mathbf{Y}$  and  $\mathbf{V}$ , containing the normalized flux and noise variance vectors, respectively. For qso  $i$ , we will write  $\mathbf{y}_i$  and  $\mathbf{v}_i$  to represent the corresponding observed flux and noise variance vectors, and will define  $\mathbf{V}_i = \text{diag } \mathbf{v}_i$ .

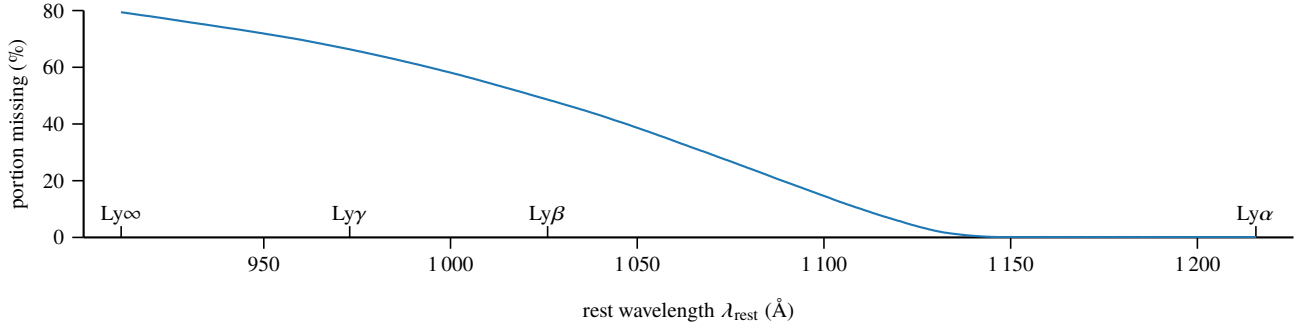
Due to masked pixels and varying redshifts of each qso, the  $\mathbf{Y}$  and  $\mathbf{V}$  matrices contain numerous missing values, especially on the blue end. Figure 2 shows the portion of available data as a function of wavelength.

#### 5.3.1 Learning the mean

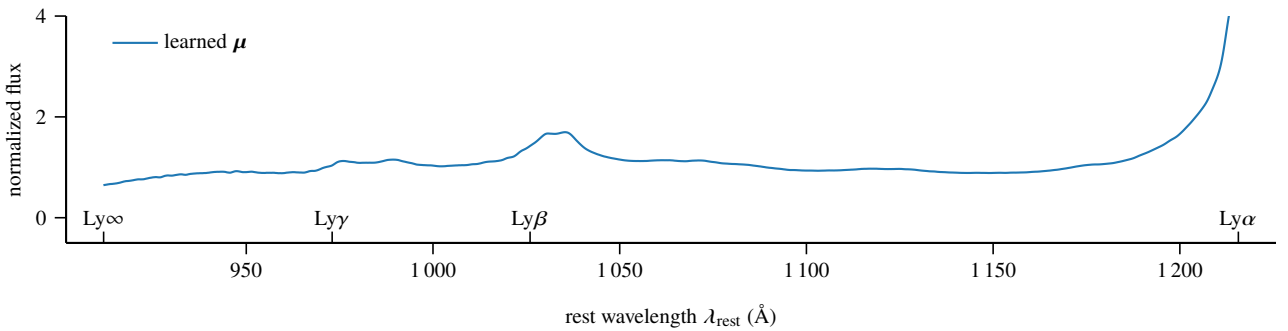
Identifying an appropriate mean vector  $\mu$  is straightforward with so many example spectra. We simply found the mean recorded value for each rest wavelength in our grid across the available measurements:

$$\mu_j = \frac{1}{N_{\text{NaN}} \sum_{y_{ij} \neq \text{NaN}}} y_{ij}. \quad (22)$$

Note that the sample mean is the maximum-likelihood estimator for  $\mu$ . The learned mean vector  $\mu$  is plotted in Figure 3. Several emission features are obvious.



**Figure 2.** The portion of missing pixels as a function of wavelength for the 48 614 qsos in the BOSS DR9 Lyman- $\alpha$  forest sample used for learning our GP model.



**Figure 3.** The learned mean vector  $\mu$  derived by taking the median across the stacked spectra. The vector has been smoothed with a 4-pixel ( $1 \text{ \AA}$ ) boxcar function for clarity on the blue end.

### 5.3.2 Learning the flux covariance and additional absorption noise

We will use standard unconstrained optimization techniques to learn the covariance matrix  $\mathbf{K}$  and absorption “noise” vector  $\omega$ . Without further structural assumptions on  $\mathbf{K}$ , however, we would be forced to learn  $N_{\text{pixels}}^2 \approx 1.5 \times 10^6$  entries. Instead we will use a low-rank decomposition to limit the number of free variables in our model:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^T, \quad (23)$$

where  $\mathbf{M}$  is an  $(N_{\text{pixels}} \times k)$  matrix with  $k \ll N_{\text{pixels}}$ . This decomposition guarantees that  $\mathbf{K}$  will be positive semidefinite (and thus a valid covariance matrix) for any  $\mathbf{M}$ . Note that this decomposition is similar to that encountered in principal component analysis (PCA); however, note that we do not constrain the columns of  $\mathbf{M}$  (the “eigenspectra”) to be orthogonal. Here we took  $k = 20$ .

We assume that each of our measured flux vectors is an independent realization drawn from a common observation prior (21):

$$p(\mathbf{Y} | \lambda, \mathbf{V}, \mathbf{M}, \omega, \mathbf{z}_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) = \prod_{i=1}^{N_{\text{spec}}} \mathcal{N}(\mathbf{y}_i; \mu, \mathbf{K} + \mathbf{\Omega} + \mathbf{V}_i), \quad (24)$$

where all NaN values are ignored. That is, in the  $i$ th entry of the product, we only use the entries of  $\mu$ ,  $\mathbf{v}_i$ , and  $\omega$ , and only the rows of  $\mathbf{M}$ , corresponding to non-masked values in  $\mathbf{y}_i$ .

We define the log likelihood of the data,  $\mathcal{L}$ , as a function of the covariance parameters  $\mathbf{M}$  and  $\omega$ . To simplify the notation, we first define the following quantities:

$$\Sigma_i = \mathbf{K} + \mathbf{\Omega} + \mathbf{V}_i; \quad (25)$$

$$\alpha_i = \Sigma_i^{-1}(\mathbf{y}_i - \mu). \quad (26)$$

Now the log likelihood is

$$\begin{aligned} \mathcal{L}(\mathbf{M}, \omega) &= \log p(\mathbf{Y} | \lambda, \mathbf{V}, \mathbf{M}, \omega, \mathbf{z}_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) \\ &= \sum_{i=1}^{N_{\text{spec}}} \log \mathcal{N}(\mathbf{y}_i; \mu, \Sigma_i) \\ &= \sum_{i=1}^{N_{\text{spec}}} -\frac{1}{2}(\alpha_i^T (\mathbf{y}_i - \mu) + \log \det \Sigma_i + N_i \log 2\pi), \end{aligned} \quad (27)$$

where  $N_i$  is the number of non-NaN pixels in  $\mathbf{y}_i$ . We will maximize  $\mathcal{L}(\mathbf{M}, \omega)$  with respect to the covariance parameters to derive our model, giving the emission model most likely to have generated our data. To enable unconstrained optimization, we parameterize the  $\omega$  parameter by its natural logarithm, guaranteeing every entry of  $\omega$  is positive after exponentiation. In the context of its role in our model, this is equivalent to reasoning about the absorption cross section  $\tau$  rather than the absorption  $\exp(-\tau)$ .

An important feature of our particular choice of model is that we can compute the matrix inverse and the log determinant of  $(\mathbf{K} + \mathbf{\Omega} + \mathbf{V})$  quickly. Namely, this matrix has the form  $\mathbf{M}\mathbf{M}^T + \mathbf{D}$ , where  $\mathbf{D}$  is diagonal. We may apply the Woodbury identity to derive

$$(\mathbf{M}\mathbf{M}^T + \mathbf{D})^{-1} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{M}(\mathbf{I} + \mathbf{M}^T\mathbf{D}^{-1}\mathbf{M})^{-1}\mathbf{M}^T\mathbf{D}^{-1}, \quad (28)$$

where the nominally  $N_{\text{pixels}} \times N_{\text{pixels}}$  inverse can be computed via a much less expensive  $k \times k$  inverse. Similarly, we may use the Sylvester determinant theorem to derive

$$\log \det(\mathbf{M}\mathbf{M}^T + \mathbf{D}) = \log \det \mathbf{D} + \log \det(\mathbf{I} + \mathbf{M}^T\mathbf{D}^{-1}\mathbf{M}), \quad (29)$$

again reducing the problem to a determinant on a  $k \times k$  matrix.

To maximize our joint log likelihood, we applied the L-BFGS algorithm, a quasi-Newton algorithm for unconstrained optimization.

The required partial derivatives are:

$$\frac{\partial \mathcal{L}_i}{\partial \mathbf{M}} = (\alpha_i \alpha_i^\top - \Sigma_i^{-1}) \mathbf{M}; \quad (30)$$

$$\frac{\partial \mathcal{L}_i}{\partial \log \omega} = \omega \circ (\alpha_i^2 - \text{diag } \Sigma_i^{-1}), \quad (31)$$

where  $\circ$  is the Hadamard (elementwise) product.

We learned the decomposed covariance matrix  $\mathbf{M}$  and  $\omega$  via L-BFGS on the selected training spectra. For this model learning phase only, we masked all pixels with noise variance larger than unity after normalization (that is, pixels with signal to noise ratios below approximately 1). Note that these pixels were only masked here and at no other point in this study. The initial value for  $\mathbf{M}$  was taken to be the top-20 principal components of  $\mathbf{Y}$ , estimated entrywise using available data. Masking low-SNR pixels was required here because PCA, in its most basic form, does not account for noise in measured values, and our heteroskedastic noise is especially troublesome. The initial value of each entry in  $\omega$  was taken to be the sample variance of the corresponding column of  $\mathbf{Y}$ , ignoring NaNs.

The first five columns of the learned  $\mathbf{M}$  and the learned absorption noise vector  $\omega$  are shown in Figure 4. The corresponding covariance matrix  $\mathbf{M}\mathbf{M}^\top$  is shown in Figure 5. Features corresponding to the Lyman series are clearly visible, including strong off-diagonal correlations between pairs of emission lines. At least seven members of the Lyman series can be identified in the covariance entries corresponding to Lyman- $\alpha$  emission. This complex (and physically correct) structure was automatically learned from the data.

We have now fully specified our GP prior for QSO emission spectra in the range  $\lambda \in [911.75 \text{ \AA}, 1216.75 \text{ \AA}]$ . Figure 6 demonstrates our model by showing an example sample from the prior distribution on QSO continua  $\mathbf{f}$ , as well as a corresponding sample from the prior distribution on observations  $\mathbf{y}$  incorporating our absorption “noise” vector  $\omega$ . The samples closely resemble actual observations.

Note that to apply our model to observations corresponding to a set of input wavelengths differing from the grid we used to learn the model, we simply interpolate (linearly) the learned  $\mu$ ,  $\mathbf{K}$ , and  $\omega$  onto the desired wavelengths. We may also account for redshift trivially should we wish to work with observed rather than rest wavelengths.

#### 5.4 Model evidence

We note that our null model  $\mathcal{M}_{\text{DLA}}$  has no parameters. Consider a set of observations of a QSO  $\mathcal{D} = (\lambda, \mathbf{y})$  with known observation noise variance vector  $\mathbf{v}$ . The model evidence for  $\mathcal{M}_{\text{DLA}}$  given by observations can be computed directly:

$$p(\mathcal{D} | \mathcal{M}_{\text{DLA}}, \mathbf{v}, z_{\text{QSO}}) \propto p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}). \quad (32)$$

The constant of proportionality is  $p(\lambda | \mathcal{M}_{\text{DLA}})$ , a quantity that we do not model here. Rather, we will assume that  $p(\lambda | \mathcal{M})$  is constant across models, causing it to cancel during the calculation of the model posterior. Therefore for the purposes of model comparison, we need only compute

$$p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = \mathcal{N}(\mathbf{y}; \mu, \mathbf{K} + \mathbf{\Omega} + \mathbf{V}), \quad (33)$$

where the  $\mu$ ,  $\mathbf{K}$ , and  $\omega$  learned above have been interpolated onto  $\lambda$ .

### 6 A GP MODEL FOR QSO SPECTRAL SIGHTLINES WITH INTERVENING DLAS

In the previous section, we learned an appropriate GP model for QSO spectra without intervening DLAS, forming our null model  $\mathcal{M}_{\text{DLA}}$ .

Here we will extend that model to create a model for sightlines containing intervening DLAS. We will first fully describe the model for spectra containing exactly one intervening DLA, then extend this model to the case of two-or-more DLAS along a line of sight. We will call our model for lines of sight containing exactly  $k$  intervening DLAS  $\mathcal{M}_{\text{DLA}(k)}$ ; here we describe  $\mathcal{M}_{\text{DLA}(1)}$ . Taking the conjunction of these models  $\{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^{\infty}$  gives our complete DLA model  $\mathcal{M}_{\text{DLA}}$ .

Consider a quasar with redshift  $z_{\text{QSO}}$ , and suppose that there is an intervening DLA along the line of sight with redshift  $z_{\text{DLA}}$  and column density  $N_{\text{H}}$ . The effect of this on our observations is to multiply the emitted flux  $f(\lambda)$  by an appropriate absorption function:

$$y(\lambda) = f(\lambda) \exp(-\tau(\lambda; z_{\text{DLA}}, N_{\text{H}})) + \varepsilon, \quad (34)$$

where  $\varepsilon$  is additive Gaussian noise due to measurement error and  $\tau$  is the absorption cross section, which has a contribution corresponding to each transition we wish to model. Here we model absorption for several members of the Lyman series:

$$\tau(\lambda; z_{\text{DLA}}, N_{\text{H}}) = N_{\text{H}} \frac{\pi e^2 f \lambda'}{m_e c} \phi(v, \sigma, \gamma), \quad (35)$$

where  $e$  is the elementary charge,  $\lambda'$  is the transition wavelength ( $\lambda' = 1215.6701 \text{ \AA}$  for Lyman- $\alpha$ ), and  $f$  is the oscillator strength of the transition ( $f = 0.4164$  for Lyman- $\alpha$ ). The line profile function  $\phi$  is a Voigt profile, where  $v$  is the relative velocity:

$$v = c \left( \frac{\lambda}{\lambda' (1 + z_{\text{DLA}})} - 1 \right), \quad (36)$$

$\sigma$  is the standard deviation of the Gaussian (Maxwellian) broadening contribution:

$$\sigma = \sqrt{\frac{kT}{m_p}}, \quad (37)$$

and  $\gamma$  is the width of the Lorentzian broadening contribution:

$$\gamma = \frac{\Gamma \lambda'}{4\pi}, \quad (38)$$

where  $\Gamma$  is the transition rate ( $\Gamma = 6.265 \times 10^8 \text{ s}^{-1}$  for Lyman- $\alpha$ ). Here we fixed the gas temperature  $T$  to  $10^4 \text{ K}$ . This has little effect, as the DLA profile is dominated by Lorentzian damping wings. Here we considered line profiles corresponding to Lyman- $\alpha$ ,  $-\beta$ , and  $-\gamma$  absorption, which we may compute for a given set of wavelengths given the known transition parameters, the gas temperature  $T$ , as well as  $z_{\text{DLA}}$  and  $N_{\text{H}}$ .

Thankfully, Gaussian processes provide a simple mechanism to model the multiplicative effect introduced by the absorption function  $\exp(-\tau)$ . Suppose that a function  $f$  has a Gaussian process prior distribution:

$$p(f) = \mathcal{GP}(f; \mu, K), \quad (39)$$

and let  $a(\lambda)$  be a known function. Then the distribution of the product  $g(\lambda) = a(\lambda)f(\lambda)$  is also a Gaussian process (GPs are closed under affine transformations):

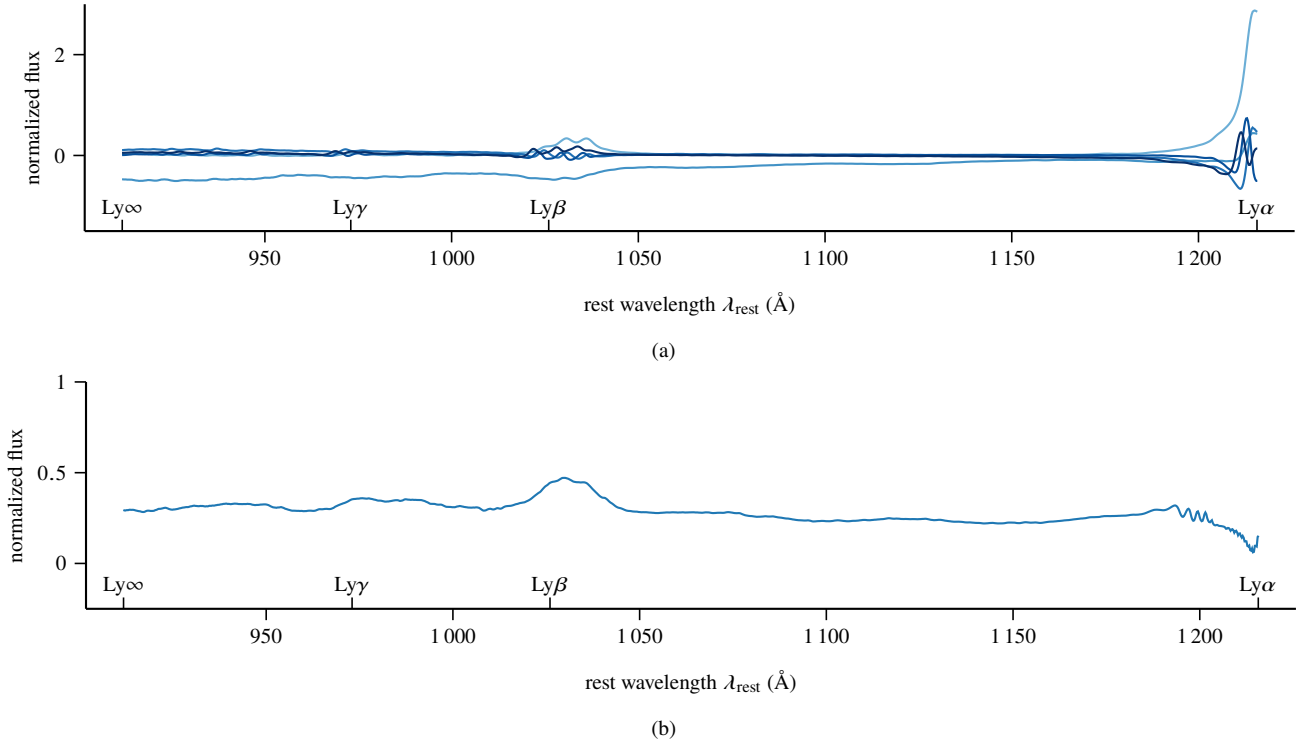
$$p(g) = \mathcal{GP}(g; \mu', K'), \quad (40)$$

where

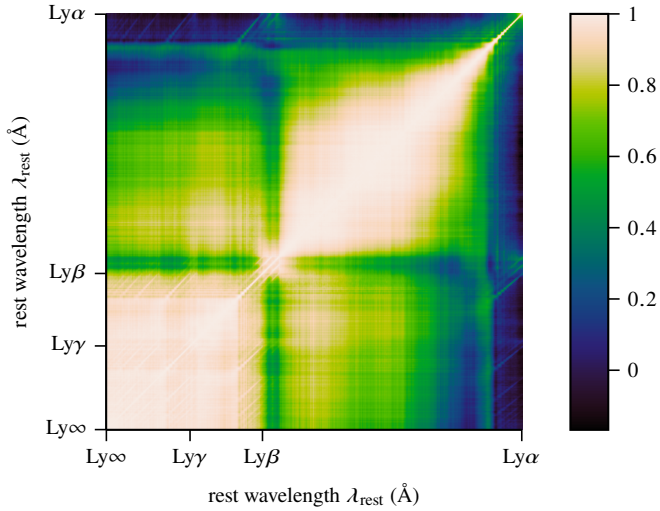
$$\mu'(\lambda) = a(\lambda)\mu(\lambda); \quad (41)$$

$$K'(\lambda, \lambda') = a(\lambda)K(\lambda, \lambda')a(\lambda'). \quad (42)$$

Therefore, given the parameters  $(z_{\text{DLA}}, N_{\text{H}})$  of a putative DLA, we compute the appropriate absorption function  $\exp(-\tau(\lambda; z_{\text{DLA}}, N_{\text{H}}))$  and modify the null GP model from the previous section as above.



**Figure 4.** (a): The first five columns of the learned  $\mathbf{M}$  and (b): the learned absorption noise vector  $\omega$ , both learned from the 48 614 qsos in the BOSS DR9 Lyman- $\alpha$  forest sample. Both have been smoothed with a 4-pixel (1 Å) boxcar function for clarity on the blue end.



**Figure 5.** The observation covariance matrix  $\mathbf{K}$  corresponding to the learned parameters shown in Figure 4. The entries have been normalized to give unit diagonal; the entries are therefore correlations rather than raw covariances.

Specifically, consider observations of a qso sightline at rest wavelengths  $\lambda$ . Our model for the corresponding emitted flux  $\mathbf{f}$  remains

$$p(\mathbf{f} | \lambda, z_{\text{qso}}) = \mathcal{N}(\mathbf{f}; \mu, \mathbf{K}). \quad (43)$$

Given the observation noise variance vector  $\mathbf{v}$ , the prior for the observation vector  $\mathbf{y}$  without intervening DLAs is

$$p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{qso}}, \mathcal{M}_{\text{DLA}}) = \mathcal{N}(\mathbf{y}; \mu, \mathbf{K} + \mathbf{\Omega} + \mathbf{V}). \quad (44)$$

Suppose now that we wish to model the observed flux with a DLA at

known redshift  $z_{\text{DLA}}$  and column density  $N_{\text{HI}}$ . First we compute the theoretical absorption function with these parameters at  $\lambda$ ; call this vector  $\mathbf{a}$ :

$$\mathbf{a} = \exp(-\tau(\lambda; z_{\text{DLA}}, N_{\text{HI}})). \quad (45)$$

Now, applying the result above, the prior for  $\mathbf{y}$  with the specified DLA is

$$p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{qso}}, z_{\text{DLA}}, N_{\text{HI}}, \mathcal{M}_{\text{DLA}(1)}) = \mathcal{N}(\mathbf{y}; \mathbf{a} \circ \mu, \mathbf{A}(\mathbf{K} + \mathbf{\Omega})\mathbf{A} + \mathbf{V}), \quad (46)$$

where  $\mathbf{a} = \text{diag } \mathbf{A}$ .

Figure 7 displays a draw from our DLA prior corresponding to the null model sample in Figure 6. Again, the sample from our DLA model  $\mathcal{M}_{\text{DLA}}$  resembles observed spectra closely.

An important feature of this model is that it is not in any way specific to DLAs. Our GP model for quasar emission spectra could be modified in an identical manner to model observed flux associated with any desired absorption feature.

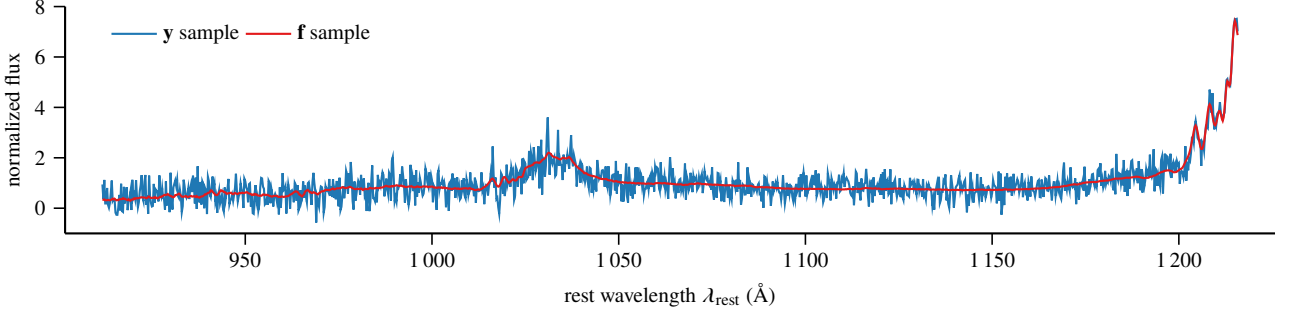
## 6.1 Model evidence

Unlike our null model, which was parameter free, our DLA model  $\mathcal{M}_{\text{DLA}(1)}$  contains two parameters describing a putative DLA: the redshift  $z_{\text{DLA}}$  and column density  $N_{\text{HI}}$ . We will denote the model parameter vector by  $\theta = (z_{\text{qso}}, N_{\text{HI}})$ . To compute the model evidence, we must compute the following integral:

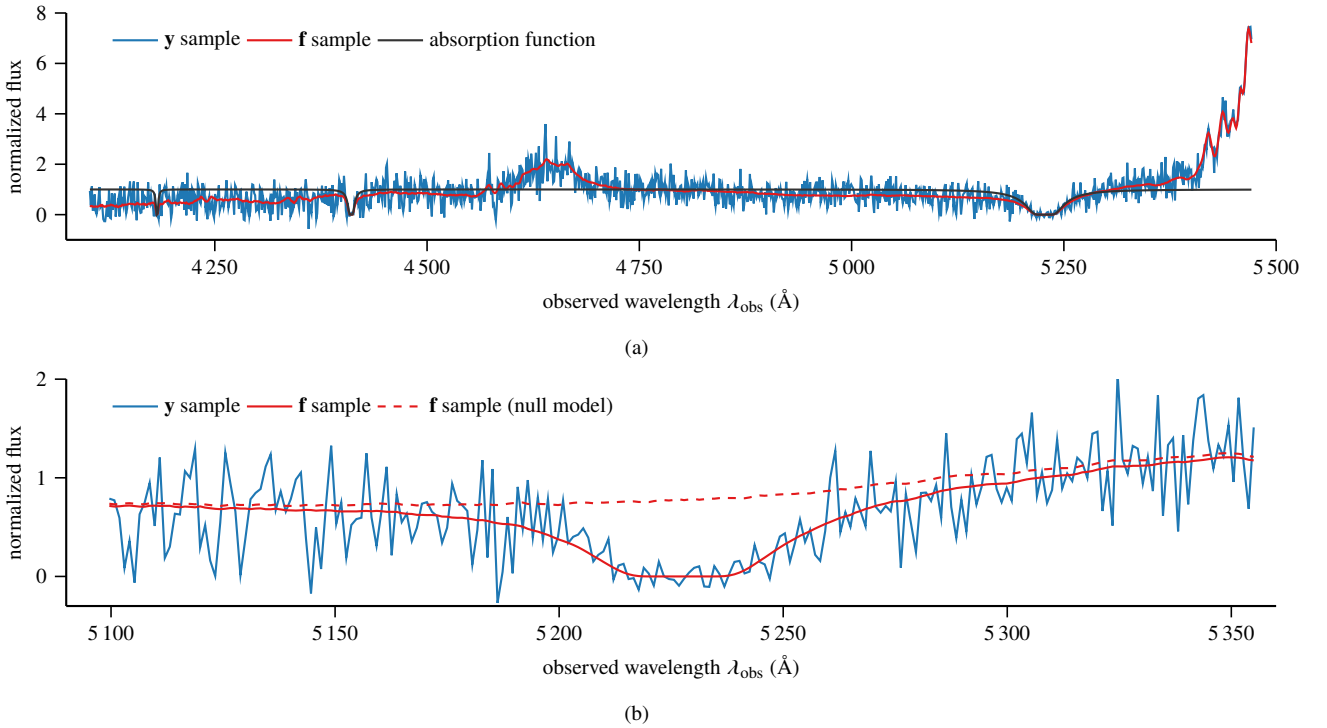
$$p(\mathcal{D} | \mathcal{M}_{\text{DLA}(1)}, \mathbf{v}, z_{\text{qso}}) \propto p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{qso}}, \mathcal{M}_{\text{DLA}(1)}) = \int p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{qso}}, \theta, \mathcal{M}_{\text{DLA}(1)}) p(\theta | z_{\text{qso}}, \mathcal{M}_{\text{DLA}(1)}) d\theta, \quad (47)$$

where we have marginalized the parameters given a prior distribution  $p(\theta | z_{\text{qso}}, \mathcal{M}_{\text{DLA}(1)})$ . Before we describe the approximation of this





**Figure 6.** An example sample from our learned qso emission spectrum model  $\mathcal{GP}(\mathbf{f}; \boldsymbol{\mu}, \mathbf{K})$  (in red), and the corresponding sample after incorporating our additional absorption correction into the model, a draw from  $p(\mathbf{y} | \boldsymbol{\lambda}, \mathbf{v}, \mathcal{M}_{\text{DLA}}) = \mathcal{GP}(\mathbf{y}; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V})$  (in blue). Constant observation noise with variance  $\nu = 0.1^2$  was simulated for the  $\mathbf{y}$  sample.



**Figure 7.** An example sample from our model for qso emission spectra with one DLA along the line of sight. Here we simulate a qso with  $z_{\text{QSO}} = 2.5$  with a DLA at  $z_{\text{DLA}} = 2.2$  and  $\log_{10} N_{\text{HI}} = 20.8$ . This sample corresponds to that in Figure 6, but is instead drawn from the DLA model with the appropriate absorption profile (plotted in grey). In (a), we show the entire simulated observations, and in (b) we show detail in the region of the Lyman- $\alpha$  absorption central wavelength, with the continuum sample from Figure 6 for comparison. Note that the full sample also reflects corresponding Lyman- $\beta$  and Lyman- $\gamma$  absorption.

integral, we will first describe the prior distribution used in our experiments.

## 6.2 Parameter prior

First, we make the assumption that absorber redshift and column density are conditionally independent given  $z_{\text{QSO}}$  and that the column density is independent of the qso redshift  $z_{\text{QSO}}$ :

$$p(\theta | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) = p(z_{\text{DLA}} | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})p(N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)}). \quad (48)$$

For the distribution  $p(z_{\text{DLA}} | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$ , we define the following range of allowable  $z_{\text{DLA}}$ :

$$z_{\text{min}} = \max \left\{ \frac{\lambda_{\text{Ly}\alpha} (1 + z_{\text{QSO}}) - 1 + 3000 \text{ km s}^{-1}/c}{\frac{\lambda_{\text{Ly}\alpha}}{\min \lambda_{\text{obs}}} - 1} \right\} \quad (49)$$

$$z_{\text{max}} = z_{\text{QSO}} - 3000 \text{ km s}^{-1}/c; \quad (50)$$

that is, we insist the absorber center be within the range of observed wavelengths (after restricting to  $\lambda_{\text{rest}} \in [911.75 \text{ Å}, 1216.75 \text{ Å}]$ ). We also apply a conservative cutoff of  $3000 \text{ km s}^{-1}$  in the immediate vicinity of the qso to avoid proximity ionization effects, and in the immediate vicinity of the Lyman limit in the quasar restframe (if visible) to avoid problems caused by possible incorrect determination of  $z_{\text{QSO}}$ .

Given these, we simply take a uniform prior distribution on

this range:

$$p(z_{\text{DLA}} | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) = \mathcal{U}[z_{\text{min}}, z_{\text{max}}]. \quad (51)$$

The column density prior  $p(N_{\text{HI}} | \mathcal{M}_{\text{DLA}})$  is slightly more complicated. We first make a nonparametric estimate of the density given the examples contained in the DLA catalog provided with the BOSS DR9 Lyman- $\alpha$  forest sample. Due to the large dynamic range of column densities, we instead choose a prior on its base-10 logarithm,  $\log_{10} N_{\text{HI}}$ .

We use the reported  $\log_{10} N_{\text{HI}}$  values for the  $N_{\text{DLA}} = 5854$  DLAs contained in the DR9 sample to make a kernel density estimate of the density  $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$ . Kernel density estimation entails centering small so-called “kernel” functions on each observation and summing them to form our estimate. Here we selected the univariate Gaussian probability density function for our kernels, with bandwidth selected via a plug-in estimator that is optimal for normal densities. The final estimate is:

$$p_{\text{KDE}}(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)}) = \frac{1}{N_{\text{DLA}}} \sum_{i=1}^{N_{\text{DLA}}} \mathcal{N}(\log_{10} N_{\text{HI}}; \ell_i, \sigma^2), \quad (52)$$

where  $\ell_i$  is the base-10 logarithm of the  $i$ th observed column density. To account for some possible systematic bias in estimating this distribution, we make two adjustments. First, we simplify the form of the distribution by fitting a parametric prior to the nonparametric kernel density estimate of the form

$$p_{\text{KDE}}(\log_{10} N_{\text{HI}} = N | \mathcal{M}_{\text{DLA}(1)}) \approx q(\log_{10} N_{\text{HI}} = N) \propto \exp(aN^2 + bN + c); \quad (53)$$

the values we learned, via least-squares fitting to the log probability over the range  $\log_{10} N_{\text{HI}} \in [20, 22]$ , were

$$a = -1.2695 \quad b = 50.863 \quad c = -509.33. \quad (54)$$

Finally, to account for some possible observation bias in the concordance catalog, we take a mixture of this this approximate column density prior with a simple log-uniform prior over a wide dynamic range:

$$p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)}) = \alpha q(\log_{10} N_{\text{HI}} = N) + (1 - \alpha) \mathcal{U}[20, 23]. \quad (55)$$

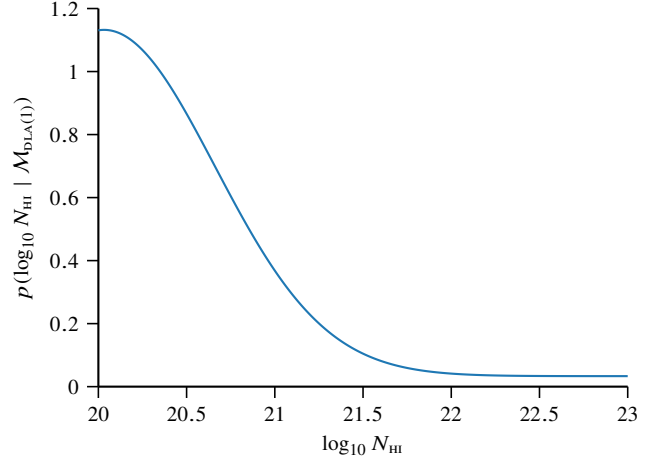
Here the mixture coefficient  $\alpha = 0.9$  favors the data-driven prior. The upper limit of  $\log_{10} N_{\text{HI}} = 23$  is more than sufficient to model all thus-far observed DLAs. The final prior  $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$  is shown in Figure 8, showing the expected bias towards smaller column densities.

### 6.3 Approximating the model evidence

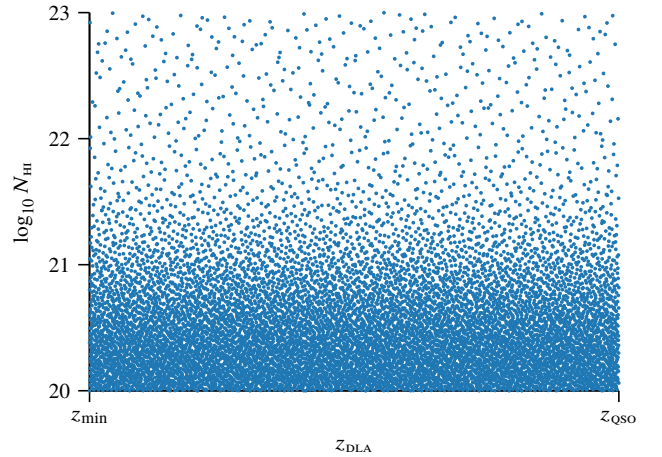
Given our choice of parameter prior, the integral in (47) is unfortunately intractable, so we will resort to numerical integration. In particular, we will use quasi-Monte Carlo (QMC) integration (Caffisch 1998). In QMC, we select  $N$  parameter samples  $\{\theta_i\}$ , evaluate the model likelihood given each of these samples, and approximate the integral in (47) by the sample mean:

$$p(\mathbf{y} | \boldsymbol{\lambda}, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) \approx \frac{1}{N} \sum_{i=1}^N p(\mathbf{y} | \boldsymbol{\lambda}, \mathbf{v}, z_{\text{QSO}}, \theta_i, \mathcal{M}_{\text{DLA}(1)}). \quad (56)$$

This is the same estimator encountered in standard Monte Carlo integration, which selects the samples by sampling independently from the parameter prior  $p(\theta | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$ . Quasi-Monte Carlo differs from normal Monte Carlo integration in that the samples  $\{\theta_i\}$  are taken from a so-called *low-discrepancy sequence*, which



**Figure 8.** The probability density function of the log column density prior used in the experiments,  $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$ .



**Figure 9.** The 10000 samples of  $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$  used to estimate the model evidence of  $\mathcal{M}_{\text{DLA}(1)}$  for a single DLA. Note that  $z_{\text{min}}$  and  $z_{\text{max}}$  are variable for each spectrum.

guarantees the chosen samples are evenly distributed, leading to faster convergence. Here we used  $N = 10000$  samples generated from a scrambled Halton sequence (Kocis & Whiten 1997) to define our parameter samples. Note that the Halton sequence gives values approximately uniformly distributed on the unit square  $[0, 1]^2$ , which (after a trivial transformation) agrees in density with our redshift prior  $p(z_{\text{DLA}} | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$ , but not our column density prior  $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$ . To correct for this, we used inverse transform sampling to endow the generated samples with the appropriate distribution. For the inverse transformation, we used the approximated inverse cumulative distribution function corresponding to our prior in (55). The generated parameter samples  $\{\theta_i\}$  are plotted in Figure 9.

Note that we can use the same technique to approximate other quantities of interest. For example, if we wish to restrict our search to only DLAs with a certain minimum column density (for example,  $\log_{10} N_{\text{HI}} > 22$ ), we can simply discard all parameter samples out of

range, giving an unbiased estimate of the desired integral:

$$\int_{z_{\min}}^{z_{\max}} \int_{22}^{\infty} p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \theta, \mathcal{M}_{\text{DLA}(1)}) p(\theta | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) dz_{\text{DLA}} d\log_{10} N_{\text{HI}}. \quad (57)$$

Note such estimators will, however, have higher variance due to the discarded parameter samples.

#### 6.4 Multiple DLAs

While the catalog we produce considers only one DLA per sightline, our model for qso sightlines containing DLAs can readily model sightlines containing two or more intervening DLAs. Again, given the parameters  $(z_{\text{DLA}}, N_{\text{HI}})$  of each absorber along the line of sight, we may compute the corresponding absorption function  $a$  and compute the observation posterior as in (46).

Let  $\mathcal{M}_{\text{DLA}(k)}$  represent a model explaining exactly  $k$  DLAs along the line of sight; we described  $\mathcal{M}_{\text{DLA}(1)}$  in the preceding sections. The model evidence integral (47) for  $\mathcal{M}_{\text{DLA}(k)}$  remains the same; however  $\theta$  will have dimension  $2k$ . Furthermore, we must consider the joint parameter prior  $p(\theta | \mathcal{M}_{\text{DLA}(k)})$ .

We propose a (nearly) independent prior between each set of DLA parameters; the dependence will be discussed later. Rather than generating a  $2k$ -dimensional low-discrepancy sequence these parameters, we propose a stepwise approach. Given a spectrum, we first use the  $\mathcal{M}_{\text{DLA}(1)}$  parameter samples  $\{\theta_i\}$  described above to approximate the model evidence (47). We can then approximate the posterior distribution of the single-DLA parameters by normalization:

$$p(\theta | z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(1)}) \propto p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \theta, \mathcal{M}_{\text{DLA}(1)}). \quad (58)$$

We may decompose the  $\mathcal{M}_{\text{DLA}(2)}$  parameters as  $\theta = [\theta_1, \theta_2]^\top$ , where each  $\theta_i$  component describes a single DLA. We propose the following prior for the  $\mathcal{M}_{\text{DLA}(2)}$  model:

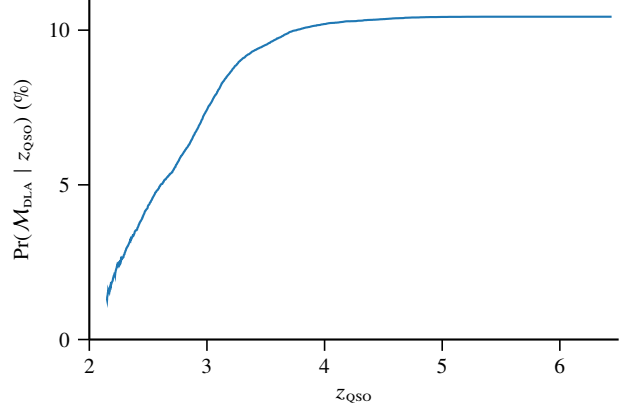
$$p(\theta_1, \theta_2 | z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(2)}) = p(\theta_1 | z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(1)}) p(\theta_2 | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}). \quad (59)$$

That is, we use the posterior probabilities from the analysis of the  $\mathcal{M}_{\text{DLA}(1)}$  model as the prior for the parameters of *one* of the DLAs when considering the  $\mathcal{M}_{\text{DLA}(2)}$  model. The prior for the parameters of the other DLA remains the noninformative prior as described above and plotted in Figure 9. For models  $\mathcal{M}_{\text{DLA}(k)}$  with  $k > 2$ , we apply a similar approach, where we combine a noninformative prior for  $\theta_k$  with an informed prior for  $\{\theta_i\}_{i=1}^{k-1}$ :

$$p(\{\theta_i\} | z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(k)}) = p(\{\theta_i\}_{i=1}^{k-1} | z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(k-1)}) p(\theta_k | z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}). \quad (60)$$

We do suggest injecting a small amount of dependence between the DLA parameters; specifically, any samples where any pair of  $z_{\text{DLA}}$  values correspond to a small relative velocity should be discarded to avoid samples describing two discrete DLAs in the same region of space.

In practice, the above scheme can be realized by first processing the spectrum with model  $\mathcal{M}_{\text{DLA}(1)}$ ; we then approximate the  $\theta_1$  posterior by renormalizing. To process the spectrum with model  $\mathcal{M}_{\text{DLA}(2)}$ , we loop through the samples in Figure 9, each sample providing  $\theta_2$ . For each sample, we sample a corresponding  $\theta_1$  sample from the approximate posterior. If the  $z_{\text{DLA}}$  values are too close, we discard the sample; otherwise, we have a valid  $\theta$  sample with which to approximate the model evidence for  $\mathcal{M}_{\text{DLA}(2)}$ . For  $\mathcal{M}_{\text{DLA}(k)}$  we proceed



**Figure 10.** The redshift-dependent model prior  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$  computed from the BOSS DR9 Lyman- $\alpha$  forest sample with parameter  $z' = 30\,000 \text{ km s}^{-1}/c$ .

in a similar way, using some minor bookkeeping to approximate the  $\{\theta_i\}_{i=1}^{k-1}$  posterior.

Note that the catalog we produce considers only  $\mathcal{M}_{\text{DLA}(1)}$ .

#### 7 MODEL PRIOR

Given a set of spectroscopic observations  $\mathcal{D}$ , our ultimate goal is to compute the probability the qso sightline contains a DLA:  $p(\mathcal{M}_{\text{DLA}} | \mathcal{D})$ . As described above, the Bayesian model selection approach requires two components: the data-independent prior probability that sightline contains a DLA,  $\Pr(\mathcal{M}_{\text{DLA}})$ , and the ability to compute the ratio of model evidences  $p(\mathcal{D} | \mathcal{M}_{\text{DLA}})$  and  $p(\mathcal{D} | \mathcal{M}_{\text{noDLA}})$ . The GP model built above allows us to compute the latter; in this section we focus on the former.

Only approximately 10% of the qso sightlines in the DR9 release contain DLAs. A simple approach to prior specification would be to use a fixed value of  $\Pr(\mathcal{M}_{\text{DLA}}) \approx 1/10$ . However, it is less likely to observe a DLA in low-redshift qsos due to the wavelength coverage of the SDSS and BOSS spectrographs being limited to  $\lambda_{\text{obs}} = 3\,800 \text{ \AA}$  and  $\lambda_{\text{obs}} = 3\,650 \text{ \AA}$ , respectively, on the blue end. Therefore, here we will use a slightly more sophisticated approach and derive a redshift-dependent prior  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$ .

Our prior will be simple and data driven. Consider a qso with redshift  $z_{\text{QSO}}$ . Let  $N$  be the number of qsos in the training sample with redshift less than  $z_{\text{QSO}} + z'$ , where  $z'$  is a small constant. Here we took  $z' = 30\,000 \text{ km s}^{-1}/c$ . Let  $M$  be the number of the sightlines of these containing DLAs. We define

$$\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}) = \frac{M}{N}, \quad (61)$$

The constant  $z'$  serves to ensure that qsos with very small redshift have sufficient data for estimating the prior. The resulting prior  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$  calculated from the DR9 sample is plotted in Figure 10.

If we wish to break down our DLA prior  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$  into its component parts, for example to find  $\Pr(\mathcal{M}_{\text{DLA}(1)} | z_{\text{QSO}})$ , we assume that DLA occurrence is independent; therefore, the probabilities multiply. If  $\frac{M}{N}$  of sightlines contain at least one DLA, then  $\frac{M^2}{N^2}$  contain at least two DLAs, etc., giving:

$$\Pr(\mathcal{M}_{\text{DLA}(k)} | z_{\text{QSO}}) = \left(\frac{M}{N}\right)^k - \left(\frac{M}{N}\right)^{k+1}. \quad (62)$$

## 8 EXAMPLE

We have now developed all of the mathematical machinery required to compute the posterior odds that a given quasar sightline contains an intervening DLA, given a set of noisy spectroscopic observations  $\mathcal{D}$ . Briefly, we summarize the steps below, using the example from Figure 1. We limit this example to searching for a single DLA, using only  $\mathcal{M}_{\text{DLA}(1)}$ .

Consider a quasar with known redshift  $z_{\text{QSO}}$ , and suppose we have made spectroscopic observations of the object  $\mathcal{D} = (\lambda, \mathbf{y})$ , with known observation noise variance vector  $\mathbf{v}$ . First, we compute the prior probability of the DLA model  $\mathcal{M}_{\text{DLA}}$ ,  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$  (61). This allows us to compute the prior odds in favor of the DLA model:

$$\frac{\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})}{\Pr(\mathcal{M}_{\neg\text{DLA}} | z_{\text{QSO}})} = \frac{\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})}{1 - \Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})}. \quad (63)$$

For our example,  $\Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}) = 10.0\%$ , giving prior odds of 0.111 (9-to-1 against the DLA model).

Next, we compute the Bayes factor in favor of the DLA model:

$$\frac{p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})}{p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}})}. \quad (64)$$

See (33) for how to compute the model likelihood for the null model and (47) for our approximation to the DLA model likelihood. For our DLA example, the Bayes factor overwhelmingly supports the DLA model, with a value of  $\exp(105) \approx 10^{45}$ . The computation of the Bayes factor is illustrated in Figure 11, which shows the prior GP mean for the null model (Figure 11a), the log likelihoods for the DLA model parameter samples (Figure 11b), and the prior GP mean for the best DLA model parameter sample (Figure 11c).

Finally, the posterior odds in favor of the sightline containing an intervening DLA is the product of (63) and (64). In practice, due to the typically large dynamic range of these quantities, it is numerically more convenient to compute the log odds. The log odds in favor of  $\mathcal{M}_{\text{DLA}}$  for the example from Figure 1 are 102 nats,<sup>8</sup> and the probability of the sightline containing a DLA is effectively unity. The DLA parameter sample with the highest likelihood was  $(z_{\text{DLA}}, \log_{10} N_{\text{H}}) = (3.285, 20.24)$ , closely matching the values reported in the DLA concordance catalog  $(z_{\text{DLA}}, \log_{10} N_{\text{H}}) = (3.283, 20.39)$ .

## 9 CATALOG

To verify the validity of our proposed method, we computed the posterior probability of  $\mathcal{M}_{\text{DLA}}$  for 162 861 quasar sightlines in the DR12Q release<sup>9</sup> of SDSS—III. The full DR12Q catalog contains 297 301 quasars, to which we applied the following cuts:

- We eliminate low-redshift ( $z_{\text{QSO}} < 2.15$ ) quasars. A total of 113 030 quasars in DR12Q satisfy this removal condition.
- We eliminate broad absorption line (BAL) quasars, determined by the BAL visual inspection survey results in the BAL\_VI field of the catalog. A total of 29 580 quasars in DR12Q satisfy this removal condition.
- We eliminate quasars that have fewer than 200 non-masked pixels in the range  $\lambda_{\text{rest}} \in [911.75, 1216.75] \text{ \AA}$ . A total of 130 quasars in DR12Q satisfy this removal condition.

<sup>8</sup> Nats are the logarithmic unit analogous to bits or dex corresponding to the base of the natural logarithm.

<sup>9</sup> <http://www.sdss.org/dr12/algorithms/boss-dr12-quasar-catalog/>

A small number of quasars satisfy multiple removal conditions.

For each of the remaining spectra, we computed the posterior probability of the  $\mathcal{M}_{\neg\text{DLA}}$  and  $\mathcal{M}_{\text{DLA}(1)}$  models, given the observations, as described in the previous sections. We produce a full catalog of our results, comprising two tables, the first rows of which are shown in Tables 1 and 2. The full catalog will be available electronically alongside this manuscript.

For each spectrum analyzed, the results catalog includes:

- SDSS metadata related to the object and spectrum,
- the range of redshifts searched for DLAs,  $[z_{\text{min}}, z_{\text{max}}]$ ,
- the log model prior,  $\log \Pr(\mathcal{M} | z_{\text{QSO}})$ , for each model considered,
- the log model evidence,  $\log p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M})$ , for each model considered,
- the model posterior,  $\Pr(\mathcal{M} | \mathcal{D}, z_{\text{QSO}})$ , and
- the MAP estimates of the  $\mathcal{M}_{\text{DLA}(1)}$  model's parameters.

### 9.1 Running time

Despite the mathematical sophistication of our method, the running time of our approach is quite manageable. Our implementation, the complete code for which will be made available along with this manuscript, is able to compute the model posterior over  $\mathcal{M}_{\neg\text{DLA}}$  and  $\mathcal{M}_{\text{DLA}(1)}$  in 0.5–2 seconds per spectrum on a standard Apple iMac desktop machine. For each spectrum we must compute 10 001 log likelihoods of the form (33) (one for (33) and 10 000 each for the  $\mathcal{M}_{\text{DLA}(1)}$  model (47)); however, the low-rank structure of our covariance allows us to compute each rapidly using the identities in (28) and (29). Our approach can easily scale to extremely large surveys and/or larger sample sizes.

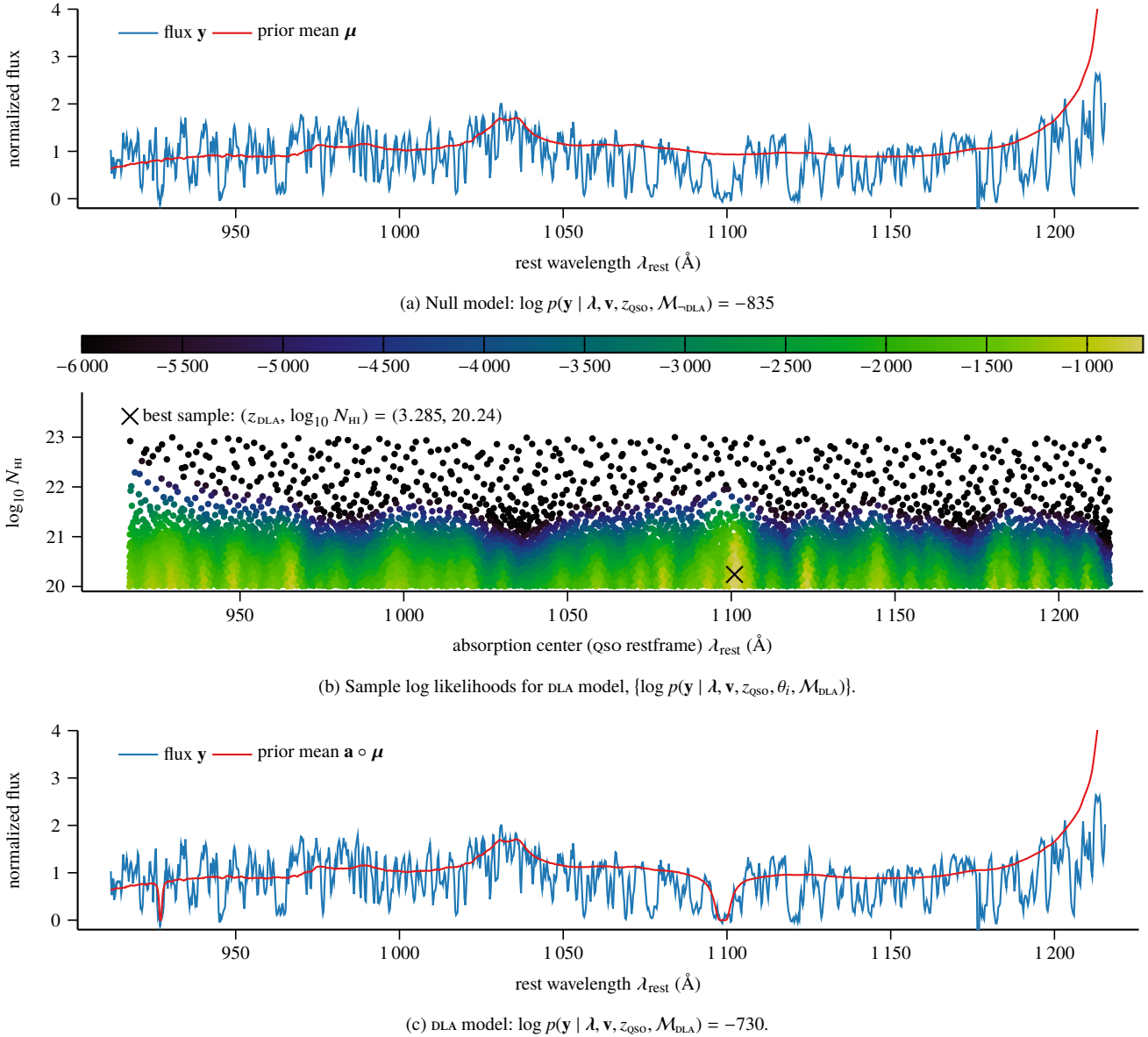
### 9.2 Analysis of results

To evaluate our results, we examined the ranking induced on the sightlines by the log posterior odds in favor of the DLA model  $\mathcal{M}_{\text{DLA}}$ . If our method is performing correctly, true DLAs should be at the top of this list, above the non-DLA-containing sightlines. To visualize the quality of our ranking, we created a receiver–operating characteristic (ROC) plot, which, for every possible threshold on the log posterior odds, plots the false positive rate (portion of non-DLAs with larger posterior odds) against the true positive rate (portion of DLAs with larger posterior odds).

Notice that creating an ROC plot requires knowledge of the ground-truth labels for each of our objects, which of course we do not have. Instead, we use the DLA concordance catalog distributed with the BOSS DR9 Lyman- $\alpha$  forest catalog as surrogate ground truth, and restrict our analysis to lines of sight that both appear in that catalog and were not removed by our cuts. A total of 54 248 objects comprise this intersection (99.7% of the catalog). The resulting ROC plot is displayed in Figure 12. The top 1%, 5%, and 10% of our ranked list, respectively, recover 44.2%, 77.3%, and 89.0% of the DLAs listed in the concordance catalog. Thus even presorting the list by the posterior probability of  $\mathcal{M}_{\text{DLA}}$  can dramatically speed up visual inspection.

A useful summary of the ROC plot is the area under the curve (AUC) statistic. The AUC has a natural interpretation: if we select a positive example and a negative example uniformly at random from those available, the AUC is the probability that the positive example would be ranked higher than the negative example. For the DR9 DLA concordance catalog surrogate, our AUC was 95.8%. Clearly our approach is effective at identifying DLAs.





**Figure 11.** An illustration of the proposed DLA-finding procedure for the quasar sightline in Figure 1. (a) shows the normalized flux with the prior GP mean for our learned null model  $\mathcal{M}_{\neg\text{DLA}}$ . (b) shows the log likelihoods for each of the parameter samples used to approximate the marginal likelihood of our DLA model  $\mathcal{M}_{\text{DLA}}$ . (c) shows the normalized flux with the prior GP mean associated with the best DLA sample,  $(z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (3.285, 20.24)$ . Notice the Lyman- $\beta$  absorption feature corresponding to this sample.

**Table 1.** The 297 301 objects in the SDSS-III DR12Q catalog, and the results of our cuts.

thing id	sdss name	plate	MJD	fiber id	right ascension	declination	$z_{\text{QSO}}$	SNR	cut flags
268514930	000000.45+174625.4	6173	56238	0528	000.0018983	+17.7737391	2.3091	000.7795	0000
(297 300 rows removed)									

An important caveat to all of the results above is that none of the surrogates is likely to represent the true ground truth, and many “false positive” sightlines could in fact contain as-yet undiscovered DLAs. Figure 13 gives an example of such a “false positive,” showing the spectrum not contained in the DLA concordance catalog that we rank the highest according to our model posterior ranking.

In fact, this spectrum appears to contain two DLAs along the line of sight. As a demonstration of our ability to detect multiple DLAs,

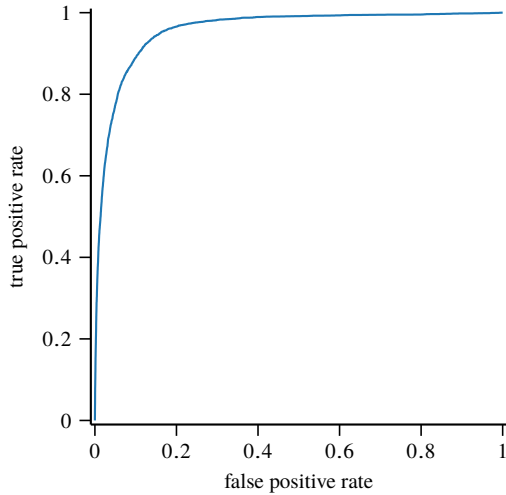
we reprocessed this spectrum using the two-DLA model  $\mathcal{M}_{\text{DLA}(2)}$ . The data overwhelmingly support  $\mathcal{M}_{\text{DLA}(2)}$  over either  $\mathcal{M}_{\text{DLA}(1)}$  or  $\mathcal{M}_{\neg\text{DLA}}$ ;  $\Pr(\mathcal{M}_{\text{DLA}(2)} \mid \mathcal{D}, z_{\text{QSO}}) = 1 - 3.1 \times 10^{-36}$ . Despite this line of sight not appearing in the DR9Q DLA concordance catalog, we do note that it was flagged during the DR12Q visual inspection.

**Table 2.** The 162 861 objects in the SDSS-III DR12Q catalog processed by our proposed GP DLA detection method, and a summary of derived quantities of interest. Note: the first nine columns match Table 1 for the included objects (those with all cut flags equal to zero).

thing id	sdss name	plate	MJD	fiber id	right ascension	declination	$z_{\text{QSO}}$	SNR	$z_{\text{min}}$	$z_{\text{max}}$
268514930	000000.45+174625.4	6173	56238	0528	000.0018983	+17.7737391	2.3091	000.7795	1.9654	2.2989
—	—	—	—	—	—	—	—	—	—	—
(162 860 rows removed)										

(cont.)	model prior		model evidence	
	$\log \Pr(\mathcal{M}_{\text{-DLA}} \mid z_{\text{QSO}})$	$\log \Pr(\mathcal{M}_{\text{DLA}} \mid z_{\text{QSO}})$	$\log p(\mathbf{y} \mid \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}})$	$\log p(\mathbf{y} \mid \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$
	-0.02915	-3.54984	-8.84158e+02	-8.83034e+02
	—	—	—	—
	(162 860 rows removed)			

(cont.)	model posterior		$\arg \max_{\theta} p(\mathbf{y} \mid \lambda, \mathbf{v}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$	
	$\Pr(\mathcal{M}_{\text{-DLA}} \mid \mathcal{D}, z_{\text{QSO}})$	$\Pr(\mathcal{M}_{\text{DLA}} \mid \mathcal{D}, z_{\text{QSO}})$	$z_{\text{DLA}}$	$\log_{10} N_{\text{HI}}$
	9.16615e-001	8.33849e-002	2.2160	20.0077
	—	—	—	—
	(162 860 rows removed)			



**Figure 12.** The ROC plot for the ranking of the 54 248 QSO sightlines contained in the BOSS DR9 Lyman- $\alpha$  forest sample (that were not filtered by our cuts), induced by the log posterior odds of containing a DLA. Ground-truth labelings were derived from the corresponding DLA concordance catalog.

### 9.3 DLA parameter estimation analysis

The main goal of our DLA-detection method is to rank QSO sightlines by their probability of containing DLAs. The computation of the evidence of our DLA model  $\mathcal{M}_{\text{DLA}}$  requires averaging over many samples of the DLA parameters ( $z_{\text{DLA}}, \log_{10} N_{\text{HI}}$ ). We may use these samples to further derive point estimates of these parameters for presumed DLAs, if desired. The simplest approach is to report the parameter sample with the highest likelihood:

$$\arg \max_{\theta_i} p(\mathbf{y} | \lambda, \mathbf{v}, z_{\text{QSO}}, \theta_i, \mathcal{M}_{\text{DLA}}); \quad (65)$$

this represents the *maximum a posteriori* (MAP) estimate of the parameters.

We analyze the behavior of the MAP estimate by comparing it with the reported values in the DR9 concordance DLA catalog.

The MAP estimates of the absorber redshift  $z_{\text{DLA}}$  are remarkably close to the catalog figures. The median difference between the two is  $-1.0 \times 10^{-4}$  ( $-30.0 \text{ km s}^{-1}$ ) and the interquartile range is  $2.2 \times 10^{-3}$

( $659 \text{ km s}^{-1}$ ). Figure 15(a) displays a kernel density estimate of the distribution of the difference between the MAP  $z_{\text{DLA}}$  estimates and the values reported in the concordance catalog, for the DLAs reported in the catalog.

The MAP estimates of the log column density  $\log_{10} N_{\text{HI}}$  show more variation with the catalog figures. The median difference between the two is quite small, only 0.026 dex. The interquartile range, however, is nontrivial at approximately 0.27 dex. Figure 15(b) displays a kernel density estimate of the distribution of the difference between the MAP  $\log_{10} N_{\text{HI}}$  values versus the values reported in the concordance catalog, for the DLAs reported in the catalog.

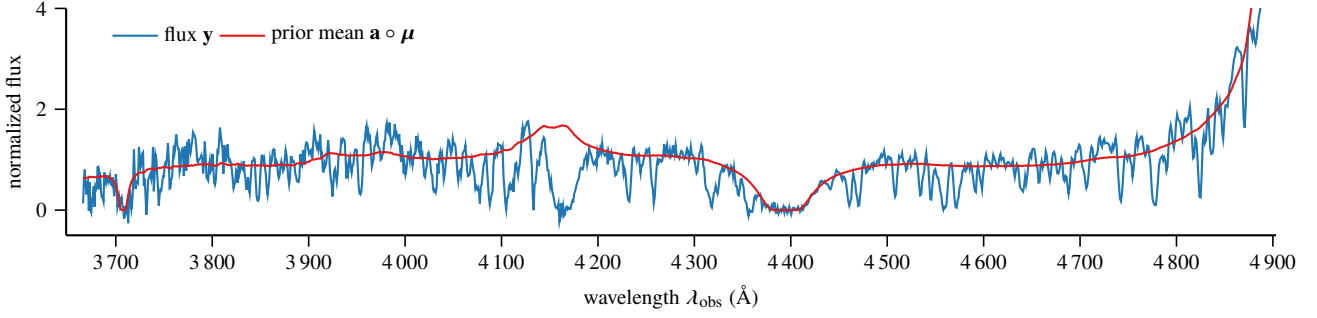
In practice, for suspected DLAs, we suggest standard procedures for Voigt-profile fitting, if an accurate estimate of the parameters is desired. Our DLA-detection procedure is primarily concerned with the evidence contained in the entire set of parameter samples, and the MAP estimate carries no special significance. In particular, several parameter ranges might have large likelihood, corresponding to several potential absorption features. The MAP estimate alone cannot convey such information.

### ACKNOWLEDGEMENTS

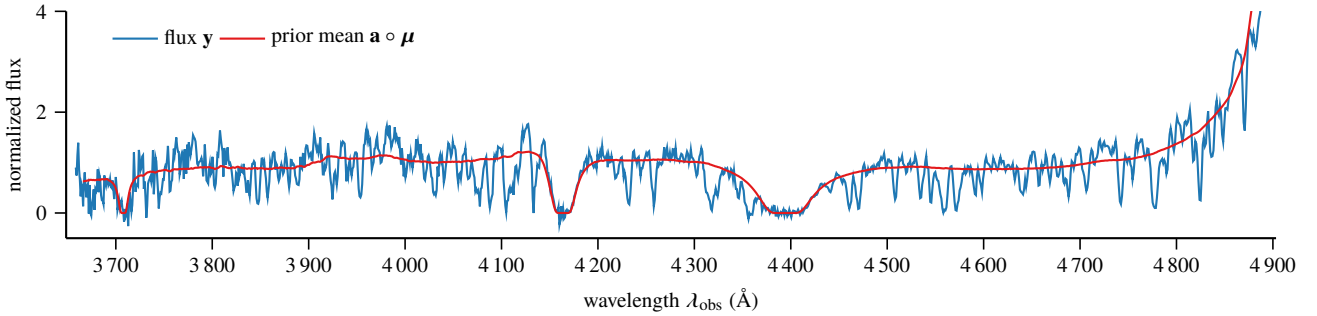
RG was supported by the National Science Foundation under Award Number IIA-1355406. SB was supported by a McWilliams Fellowship from Carnegie Mellon University and by NASA through Einstein Postdoctoral Fellowship Award Number PF5-160133.

### REFERENCES

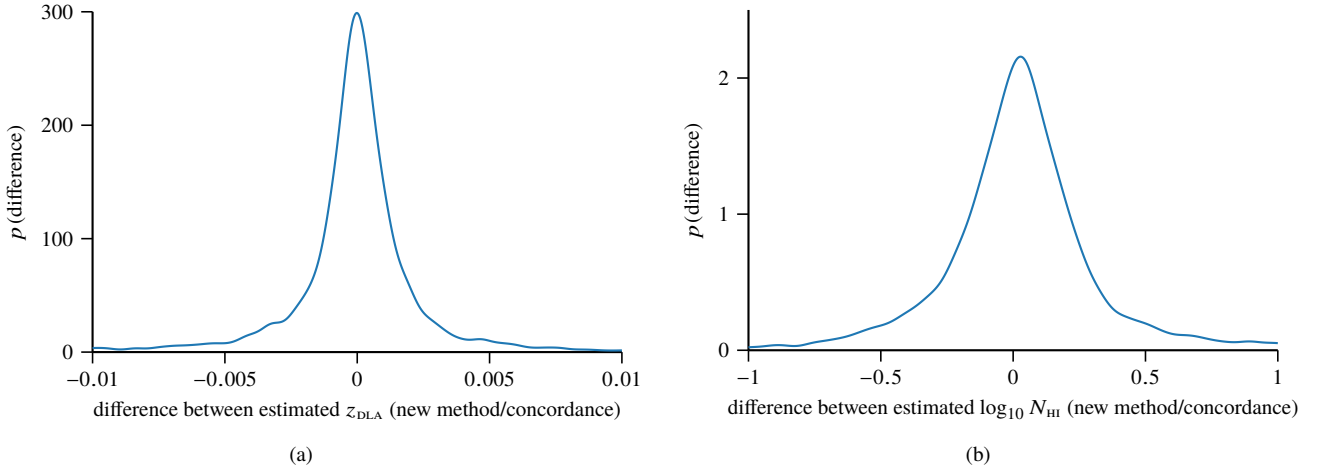
- Abazajian K. N., et al., 2009, *ApJS*, **182**, 543
- Ahn C. P., et al., 2012, *ApJS*, **203**, 21
- Ahn C. P., et al., 2014, *ApJS*, **211**, 17
- Anderson L., et al., 2012, *MNRAS*, **427**, 3435
- Anderson L., et al., 2014, *MNRAS*, **441**, 24
- Aubourg É., et al., 2015, *Phys. Rev. D*, **92**, 123516
- Bird S., Vogelsberger M., Haehnelt M., Sijacki D., Genel S., Torrey P., Springel V., Hernquist L., 2014, *MNRAS*, **445**, 2313
- Bird S., Haehnelt M., Neeleman M., Genel S., Vogelsberger M., Hernquist L., 2015, *MNRAS*, **447**, 1834
- Bovy J., et al., 2011, *ApJ*, **729**, 141



**Figure 13.** The spectrum appearing in the BOSS DR9 Lyman- $\alpha$  forest sample, not contained in the corresponding DLA concordance catalog, with the highest posterior probability of containing an DLA according to our model. The object is SDSS 235057.87-005209.9, (plate, MJD, fiber) = (4214, 55451, 52),  $z_{\text{QSO}} = 3.0207$ . We overwhelmingly believe there to be a DLA along the line of sight with most-likely parameters  $z_{\text{DLA}} = 2.6147$ ,  $\log_{10} N_{\text{HI}} = 21.194$ . The prior mean corresponding to this sample is also shown. In fact, this “false positive” appears to contain two DLAs not flagged by previous searches.



**Figure 14.** The spectrum from Figure 13, processed with the two-DLA model  $\mathcal{M}_{\text{DLA}(2)}$ . We overwhelmingly believe there to be two DLAs along the line of sight with most-likely parameters  $\theta_1 = (z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (2.6147, 21.194)$  and  $\theta_2 = (z_{\text{DLA}}, \log_{10} N_{\text{HI}}) = (2.4252, 20.648)$ . The prior mean corresponding to these parameters is also shown.



**Figure 15.** Kernel density estimate of the difference between the MAP estimates of the DLA parameters  $(z_{\text{DLA}}, \log_{10} N_{\text{HI}})$  for DLAs listed in the BOSS DR9 Lyman- $\alpha$  forest sample, against the catalog-reported values.

Caffisch R. E., 1998, *Acta Numerica*, 7, 1

Carithers W., 2012, DLA Concordance Catalog, Published internally to SDSS

Cen R., 2012, *ApJ*, 748, 121

Chen H.-W., 2005, in Braun R., ed., *Astronomical Society of the Pacific Conference Series Vol. 331, Extra-Planar Gas*. p. 371 ([arXiv:astro-ph/0410558](https://arxiv.org/abs/astro-ph/0410558))

Doi M., et al., 2010, *AJ*, 139, 1628

Eisenstein D. J., et al., 2011, *AJ*, 142, 72

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 1996, *AJ*, 111, 1748

Fumagalli M., O’Meara J. M., Prochaska J. X., Rafelski M., Kanekar N., 2015, *MNRAS*, 446, 3178

Gardner J. P., Katz N., Weinberg D. H., Hernquist L., 1997, *ApJ*, 486, 42

Gunn J. E., et al., 1998, *AJ*, 116, 3040

Gunn J. E., et al., 2006, *AJ*, 131, 2332

Haehnelt M. G., Steinmetz M., Rauch M., 1998, *ApJ*, 495, 647

Jedamzik K., Prochaska J. X., 1998, *MNRAS*, 296, 430

- Kocis L., Whiten W. J., 1997, *ACM Transactions on Mathematical Software*, 23, 266
- Le Brun V., Bergeron J., Boisse P., Deharveng J. M., 1997, *A&A*, [321](#), [733](#)
- Lee K.-G., et al., 2013, *AJ*, [145](#), [69](#)
- Maller A. H., Prochaska J. X., Somerville R. S., Primack J. R., 2001, *MNRAS*, [326](#), [1475](#)
- Noterdaeme P., et al., 2012, *A&A*, [547](#), [L1](#)
- Okoshi K., Nagashima M., 2005, *ApJ*, [623](#), [99](#)
- Pâris I., et al., 2012, *A&A*, [548](#), [A66](#)
- Pâris I., et al., 2014, *A&A*, [563](#), [A54](#)
- Pontzen A., et al., 2008, *MNRAS*, [390](#), [1349](#)
- Prochaska J. X., Wolfe A. M., 1997, *ApJ*, [487](#), [73](#)
- Prochaska J. X., Wolfe A. M., 2009, *ApJ*, [696](#), [1543](#)
- Prochaska J. X., Herbert-Fort S., Wolfe A. M., 2005, *ApJ*, [635](#), [123](#)
- Prochaska J. X., O’Meara J. M., Worseck G., 2010, *ApJ*, [718](#), [392](#)
- Rahmati A., Pawlik A. H., Raicevic M., Schaye J., 2013, *MNRAS*, [430](#), [2427](#)
- Rao S. M., Nestor D. B., Turnshek D. A., Lane W. M., Monier E. M., Bergeron J., 2003, *ApJ*, [595](#), [94](#)
- Rasmussen C. E., Williams C. K. I., 2006, *Gaussian Processes for Machine Learning*. MIT Press
- Reid B., et al., 2016, *MNRAS*, [455](#), [1553](#)
- Ross N. P., et al., 2012, *ApJS*, [199](#), [3](#)
- Slosar A., et al., 2011, *J. Cosmology Astropart. Phys.*, [9](#), [1](#)
- Smee S. A., et al., 2013, *AJ*, [146](#), [32](#)
- Smith J. A., et al., 2002, *AJ*, [123](#), [2121](#)
- Wolfe A. M., Turnshek D. A., Smith H. E., Cohen R. D., 1986, *ApJS*, [61](#), [249](#)
- Wolfe A. M., Gawiser E., Prochaska J. X., 2005, *ARA&A*, [43](#), [861](#)
- York D. G., et al., 2000, *AJ*, [120](#), [1579](#)

This paper has been typeset from a  $\text{\LaTeX}$  file prepared by the author.